Attribute Learning for Image/Video Understanding

Yanwei Fu

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2015

Abstract

For the past decade computer vision research has achieved increasing success in visual recognition including object detection and video classification. Nevertheless, these achievements still cannot meet the urgent needs of image and video understanding. The recently rapid development of social media sharing has created a huge demand for automatic media classification and annotation techniques. In particular, these types of media data usually contain very complex social activities of a group of people (e.g. YouTube video of a wedding reception) and are captured by consumer devices with poor visual quality. Thus it is extremely challenging to automatically understand such a high number of complex image and video categories, especially when these categories have never been seen before.

One way to understand categories with no or few examples is by transfer learning which transfers knowledge across related domains, tasks, or distributions. In particular, recently lifelong learning has become popular which aims at transferring information to tasks without any observed data. In computer vision, transfer learning often takes the form of attribute learning. The key underpinning idea of attribute learning is to exploit transfer learning via an intermediate-level semantic representations – attributes. The semantic attributes are most commonly used as a semantically meaningful bridge between low feature data and higher level class concepts, since they can be used both descriptively (e.g., 'has legs') and discriminatively (e.g., 'cats have it but dogs do not'). Previous works propose many different attribute learning models for image and video understanding. However, there are several intrinsic limitations and problems that exist in previous attribute learning work. Such limitations discussed in this thesis include limitations of user-defined attributes, projection domain-shift problems, prototype sparsity problems, inability to combine multiple semantic representations and noisy annotations of relative attributes. To tackle these limitations, this thesis explores attribute learning on image and video understanding from the following three aspects.

Firstly to break the limitations of user-defined attributes, a framework for learning latent attributes is present for automatic classification and annotation of unstructured group social activity in videos, which enables the tasks of attribute learning for understanding complex multimedia data with sparse and incomplete labels. We investigate the learning of latent attributes

for content-based understanding, which aims to model and predict classes and tags relevant to objects, sounds and events – anything likely to be used by humans to describe or search for media. Secondly, we propose the framework of transductive multi-view embedding hypergraph label propagation and solve three inherent limitations of most previous attribute learning work, i.e., the projection domain shift problems, the prototype sparsity problems and the inability to combine multiple semantic representations. We explore the manifold structure of the data distributions of different views projected onto the same embedding space via label propagation on a graph. Thirdly a novel framework for robust learning is presented to effectively learn relative attributes from the extremely noisy and sparse annotations. Relative attributes are increasingly learned from pairwise comparisons collected via crowdsourcing tools which are more economic and scalable than the conventional laboratory based data annotation. However, a major challenge for taking a crowdsourcing strategy is the detection and pruning of outliers. We thus propose a principled way to identify annotation outliers by formulating the relative attribute prediction task as a unified robust learning to rank problem, tackling both the outlier detection and relative attribute prediction tasks jointly.

In summary, this thesis studies and solves the key challenges and limitations of attribute learning in image/video understanding. We show the benefits of solving these challenges and limitations in our approach which thus achieves better performance than previous methods.

Declaration

I, Yanwei Fu, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date:

Some parts of the work have previously been published as:

- Chapter 3
  - Fu, Y.; Hospedales, T.; Xiang, T; Gong, S. "*Attribute Learning for Understanding Unstructured Social Activity*", European Conference on Computer Vision (ECCV) 2012;
  - Fu, Y. ; Hospedales, T. ; Xiang, T. ; Gong, S. "*Learning Multi-modal Latent Attributes*" IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI), 36(2), 303-316, Feb 2014;
- Chapter 4
  - Fu, Y.; Hospedales, T.; Xiang, T.; Fu, Z.; Gong, S. "*Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation*" European Conference on Computer

Vision (ECCV) 2014;

  - Fu, Y.; Hospedales, T.; Xiang, T.; Gong, S. "*Transductive Multi-view Zero-Shot Learning*" IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI), 2015
  - Fu, Y.;Yang, Y.; Hospedales,T.; Xiang, T.; Gong, S. "*Transductive Multi-label Zero-shot Learning*" British Machine Vision Conference (BMVC) 2014;
  - Fu, Y.;Yang, Y.; Hospedales,T.; Xiang, T.; Gong, S. "*Transductive Multi-calss and Multi-label Zero-shot Learning*" ECCV 2014 workshop on Parts and Attribute;

- Chapter 5

  - Fu, Y.; Hospedales, T.; Xiang, T.; Gong, S.; Yao, Y. "*Interestingness Prediction by Robust Learning to Rank*" European Conference on Computer Vision (ECCV) 2014;
  - Fu, Y.; Hospedales, T.; Xiang, T.; Xiong, J.; Gong, S.; Wang, Y.; Yao, Y."*Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*"submitted to IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI);

Acknowledgements

I would like to thank my supervisor, Dr. Tao Xiang, for his perpetual patience, encouragement and guidance. He has managed to keep me on track and on topic, which is a challenge as I tend to wonder off topic easily. I have learned a lot under his caring supervision. I doubt I could have had a more capable or accommodating advisor, and he has my deepest gratitude. Besides, I am deeply grateful for invaluable advice and consistent support from my co-supervisor Professor Shaogang Gong throughout my three year PhD life. I would like to thank Dr. Fabrizio Smeraldi for being my internal examiner throughout my PhD project.

I would also like to thank Dr. Timothy Hospedales. Our collaborative efforts has led to several successes. Dr. Hospedales has continuously provided stimulating conversations which has lead to an ever growing list of new ideas to try. I greatly appreciate his encouragement of trying out "dumb" algorithms.

Many thanks also to Prof. Yuan Yao and Prof. Yizhou Wang, who invited me to visit Peking University and I enjoyed a very good summer in Beijing. Without your guidance on some theories on robust ranking, I would never have completed this thesis.

I am grateful to all the students and associates at the Vision Group for their friendship and support, in particular Yi-Zhe Song, Miles Hansard, Lourdes Agapito, Xiatian Zhu (Eddy), Xun Xu (Alex), Yi Li, Hanxiao Wang, Parthipan Siva, Chen Change Loy (Cavan), Ke Chen (Cory), Zhiyuan Shi (Patrick), Ryan Layne, Zhenyong Fu (Ian), Wenzhao Li, Elyor Kodirov, Chris Russell, Nikos Pitelis, Sara Vicente, Ravi Garg, Tom Hains, Tassos Roussos, and Rui Yu. I also would like to convey my special thanks to Dr. Yanwen Guo, Dr. Feng Liu and Dr. Feng Tang who encourage me to take up the challenge of PhD study. Last but not least, I would like to thank my family for indulging my flights of fancy, like running off to UK for three years to do my PhD.

Contents

List of Figures

List of Tables

List of Acronyms

AwA Animal with Attribute

AUC Area Under Curve

BoW Bag of Words

CC Class-Conditional

DAP Direct Attribute Prediction

GF Generalised Free

M2LATM Multi-Modal Latent Attribute Topic Model

TMV-HLP Transductive Mult-view embedding Hypergraph Label Propagation

URLR Unified Robust Learning to Rank

UD User-Defined

NN Nearest-Neighbour (distance)

PCA Principle Component Analysis

SVM Support Vector Machine

ZSL Zero-Shot Learning

Chapter 1

Introduction

With the rapid developments of devices capable of digital media capture, vast volumes of multimedia data are being uploaded and shared on social media platforms (e.g. YouTube and Flickr). For example, 100 hours of video are uploaded every minute on YouTube[1]. Managing this growing volume of data demands effective techniques for automatic media understanding. Such automatic techniques are important for content-based image and video understanding in order to enable effective indexing, search, retrieval, filtering and recommendation of multimedia content from the vast quantity of image and video data.

Generally, conventional supervised learning approaches are feasible for the benchmark datasets of tens or hundreds of categories of images and videos and they work in this way (in Figure 1.1): the powerful low-level features such as SIFT [Low04], HoG [BZM07b] and recent deep features [DJV+14, KSH12, SEZ+14] are extracted from all examples, and modern machine learning classifiers such as *support vector machines* [CL01, Lam09] or *random forest* [BZM07a] are learned from an amount of well-labelled training instances. Thus to successfully identify one particular category, the supervised classifier must learn the 'knowledge' from the previous examples in the same class.

Nevertheless it is still a significant challenge to develop an automatic system for large-scale image and video understanding using a conventional supervised learning based framework (see Figure 1.1). Especially, it is very expensive and even impossible to annotate the training examples in large-scale for all real-world categories. For example, there are at least 30000 human-

[1] http://www.youtube.com/t/faq

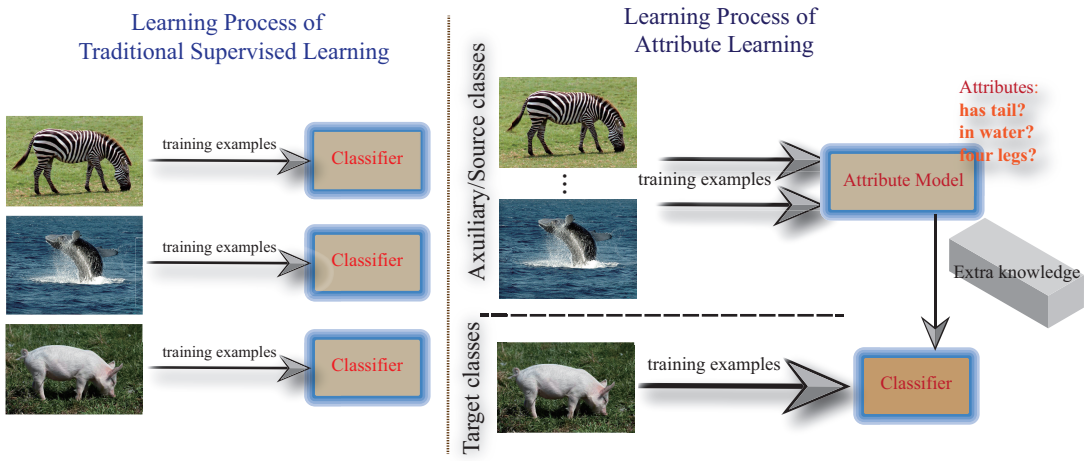


Figure 1.1: Traditional supervised learning vs. attribute learning.

distinguishable basic object classes and much more subordinate ones, e.g. different breeds of dogs [KJYFF11] and birds [WBW+11]. Humans can also create new categories dynamically from single examples or solely based on high-level description, e.g., the videos of "Germany World Cup winner celebrations 2014". Such exponential combinations of relevant categorisation tasks, if labelled and trained in the conventional supervised learning ways, needs the collection of millions or billions of training instances. Furthermore, the conventional supervised learning methods cannot identify the categories or sub-categories for which there is no positive training example. For example, the conventional supervised video classifier of "Germany World Cup winner celebrations 2014" cannot be trained until some positive video samples are available and shared after July 2014 when Germany finally won the football match over Argentina.

Different from the conventional supervised learning framework, humans have the ability of "learning to learn" [Thr96]. Specifically, we can exploit shared properties and characteristics among categories and resort to human linguistic knowledge bases in the form of written text. This inspires a recent flourish of research into attribute learning for image and video understanding. Still taking the video classification for the event "Germany World Cup winner celebrations 2014" as an example. Our strategy would be to construct the classifiers and transfer the knowledge from the previous well-labelled video examples like " FC Bayern Munich - Champions of Europe 2013" to help identify the novel unseen class. This idea inspires attribute learning for image and video understanding in this thesis, which belongs to transfer learning.

Transfer learning refers to the problem of applying the knowledge learned in one or more auxuliary tasks/domains/sources to develop an effective hypothesis for a target task/domain. In

general, transfer learning emphasizes the transfer of knowledge across domains, tasks, and distributions that are similar but not the same.

Research in transfer learning continued after 1995 under a variety of names: learning to learn, life-long learning, knowledge transfer, transfer learning, multiple task learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, meta-learning, and incremental, cumulative, and continual learning [PY10]. In this thesis, we follow Pentina et al. [PL14] and use *life-long learning* to refer to the transfer learning with the settings of transfering information to tasks for which no data have been observed so far. In general, *life-long learning* invokes some of the most important questions [**?**]:

- What is the best representation and method for transferring prior knowledge to a new task?
- How does the use of prior knowledge affect learning in the target task/domain?
- What is the nature of similarity or relatedness between tasks for the purposes of learning? Can it be measured?

As one way of implementing transfer learning, attributes are the semantic-rich representations to help bridge between low feature data and higher level class concepts. Attributes can be used both descriptively (e.g., 'has legs') and discriminatively (e.g., 'cats have it but dogs do not'). Such properties of attributes in our approach can explicitly explain to us: which knowledge to transfer; where of knowledge transfer, and why of difference knowledge sources and similarity measures for knowledge transfer [RSS$^+$10].

1.1 Attribute Learning for Image/Video Understanding

Different from the conventional supervised approaches in Figure 1.1, humans actually have very different strategies for learning: for examples, when reading '*flightless birds living almost exclusively in Antarctica*', we know it is penguin for certain even though we might have never seen a penguin in our life. In cognitive science [Thr96], studies explain that humans achieve "learning to learn" new concepts by extracting intermediate semantic representation or high-level descriptions (i.e. *flightless*, *bird*, *living in Antarctica*) and transferring knowledge from known sources (other bird classes, e.g. swan, canary, cockatoo and so on) to the unknown target (penguin). That is the reason why humans are able to understand new concepts with no or only few training samples.

Recently, inspired by "learning to learn" in humans and to minimise the necessary labelled training examples for supervised classifiers, researchers built attribute learning recognition models that are capable of classifying novel classes with no training example. In machine learning, "learning to learn" is also referring as *life-long learning*[2][PL14, Thr96, TM95]. In computer vision, the task of classifying classes without any observed data is called *zero-shot learning*. To enable such zero-shot learning algorithms, the key underpinning idea is to exploit transfer learning via an intermediate-level semantic representation (e.g. attributes) as a semantically meaningful bridge between raw data and class concepts. In particular, such an idea is realised by attribute learning [FHXG12, PP12, FEHF09, JWZ13, LNH13, LNH09, APHS13, LJLFF10, LKS11, PG11a, PG11b, KPG12, FYH+14a, FHXG13, FYH+14b, FHXG14] in the computer vision community.

The common *attribute learning pipeline* in most previous work [FHXG12, PP12, FEHF09, JWZ13, LNH13, LNH09, APHS13, LJLFF10, LKS11, PG11a, PG11b, KPG12, FYH+14a, FHXG13, FYH+14b, FHXG14] is shown in Figure 1.1. Two datasets with disjoint classes are considered: a labelled known auxiliary set (e.g. Zebra and Whale) where a semantic representation is given for each data point, and a target dataset (e.g. Pig) to be classified with no labelled instance and semantic representation. Such a semantic representation is assumed to be shared between the auxiliary and target datasets. Specifically, apart from the class label, each auxiliary data point is labelled by a semantic representation such as visual attributes [LNH09, FEHF09, LKS11, FHXG13], and other semantic representations [MCCD13, FCS+13, SGS+13, RSS12]. A projection function, mapping low-level features to the semantic space, is learned from the auxiliary dataset by either classification or regression models. Such a projection is then applied directly to map each unlabelled target class instance into the same semantic representation space. Within this semantic space, a zero-shot classifier is pre-defined by "extra-knowledge" to recognise all unseen instances. The two most popular zero-shot learning algorithms proposed in [LNH13, LNH09] are Direct Attribute Prediction (DAP) and Indirect Attribute Prediction (IAP). In particular, a single 'prototype' of each target class is specified in the semantic space. Depending on the semantic space, this prototype can be an attribute annotation vector [LNH09] or other semantic representation vector inferred from the target class name [FCS+13].

[2]it is related to *domain adaptation* which is to perform well on a new task for which only unlabeled or very few labeled data samples are observed. Life-long learning, domain adaptation and zero-shot learning all belong to transfer learning.

1.2 Challenges and Motivations

The previous attribute learning pipeline makes a good start to help solve zero-shot learning problems. It targets image and video understanding and tries to help model and predict classes and tags relevant to objects, sounds and events – anything likely to be used by humans to describe or search for media. However, there are several challenges that exist in previous attribute learning work that limits the performance on image and video understanding. This section briefly explains such challenges.

1.2.1 Limitations of User-defined Attributes

Most previous work [LNH09, FEHF09, DFV11, FZ07, LNH13, KBBN09, PH12, YCF+13, KPG12, MSN11, HSG11, FEH10, LKS11, PP12] relies on manually defined attributes. Such attributes are usually defined by users or experts or extracted from concept ontoloy, e.g. WordNet [Mil95]. These attributes are annotated at class-level or instance-level by using crowdsourcing tools. Such user-defined attributes have been used for image and video understanding.

However, these user-defined attributes are neither flexible nor comprehensive enough to explore the complex multi-modal image and video data. Specifically, user-defined attributes come from the knowledge of user experts or concept ontology and may be neither extendable nor discrimintive enough to help understand complex multi-modal[3] image and video data.

For example, consumer videos (e.g. home videos) [JYC+11] have the most common yet challenging types of video content – unstructured social group activity and these types of videos are hard to be fully explained by user-defined attributes. The main challenge comes from the unconstrained space of objects, events and interactions which make such consumer videos intrinsically very complex. Especially, this unconstrained domain gives rise to a space of possible content concepts that is orders of magnitude greater than that typically addressed by most previous video analysis work (e.g. human action recognition [LKS11]). It is thus impossible to exhaustively make user-defined attributes for all these content concepts used for zero-shot learning tasks.

To sum up, the underlying challenges of the limitations of user-defined attributes can be broadly characterised as *sparsity*, *incompleteness* and *ambiguity* of the annotations of user-defined attributes.

[3]Here, multi-modal indicates different feature types and modalities for images and videos. For example, image can be described by global features, local features, and deep features. Video content have visual, audio and action information.

Annotations of user-defined attributes are sparse. Visual data covers a huge unconstrained space of object, activity or event concepts, therefore requiring numerous tags to completely annotate the underlying content. However the number of labelled training instances per annotation of user-defined attributes is likely to be low. For example, consumer videos shared on social media platforms only have 2.5 tags on average versus 9 tags in general for YouTube videos [JYC+11]. Such tags can be taken as different types of user-defined attributes.

Annotations of user-defined attributes are intrinsically incomplete. Since the space of visual concepts is unconstrained, exhaustive manual annotation of examples for every user-defined attribute is impractically expensive, even through mechanisms such as Amazon Mechanical Turk (AMT) [SF08]. Previous studies have therefore focused on analyzing relatively constrained spaces of visual content and hence annotation ontologies [LKS11]. However, there are for example, around 30000 relevant object classes which are recognizable by humans [Bie87]. This means that any set of user-defined attributes will either be too small to provide a complete vocabulary to describe general images and videos, or have insufficient training data for every user-defined attributes.

Annotations of user-defined attributes are ambiguous. *Ambiguity* is relatively less studied in previous work but a significant challenge for image and video understanding. Even for the same image/video, subjective factors (e.g. cultural background) may lead to contradictory and ambiguous annotations. A well-known example is that some countries take a nodding head as "yes", while others as "no". This ambiguity of annotations of user-defined attributes can be taken as label noise. Ambiguity also arises from the semantic gap between annotations and raw data: semantically obvious annotations are not necessarily detectable from low-level features; while the most useful annotations for a model may not be the most semantically obvious ones that humans commonly provide. Finally, the weakly supervised nature of annotation of user-defined attributes, and the multi-modality of the data are another strong sources of ambiguity. For example, an annotation of the "clapping" attribute comes with no information detailing where it was observed (temporally) in a video, or whether it was only seen visually, only heard, or both seen and heard.

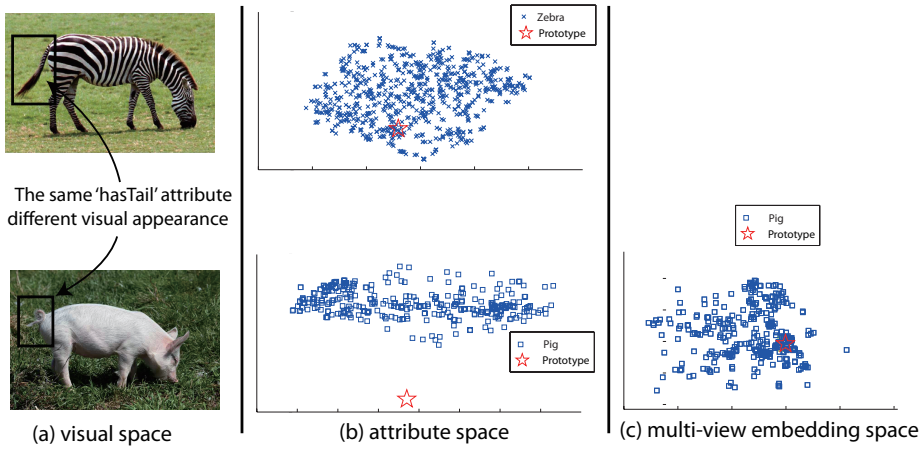


Figure 1.2: An illustration of the projection domain shift problem. Zero-shot prototypes are shown as red stars and image feature projections shown in blue.

1.2.2 Projection Domain-shift Problem

As shown in Figure 1.1, previous attribute learning pipelines can exploit knowledge transfer via an intermediate-level semantic representation which is assumed to be shared between the auxiliary/source dataset and the target/test dataset and re-used as a bridge between the source and target domains for knowledge transfer.

However, there is an inherent limitation with this pipeline. We call it *projection domain-shift problem*: Since two datasets have different and potentially unrelated classes, the underlying data distributions of the classes differ, so do the 'ideal' projection functions between the low-level feature space and the semantic spaces. Therefore, using the projection functions learned from the auxiliary dataset/domain without any adaptation to the target dataset/domain causes an unknown shift/bias.

This is further illustrated in Figure 1.2, both of Zebra (auxiliary) and Pig (target) classes in Animal with Attribute (AwA) dataset [LNH09] share the same 'hasTail' semantic attribute, yet with different visual appearance of their tails. Similarly, many other attributes of Pig are visually different from the corresponding attributes in the auxiliary classes. Figure 1.2(b) illustrates the projection domain shift problem by plotting an 85D attribute space representation of image feature projections and class prototypes: a large discrepancy between the Pig prototype and the projections of its class member instances is seen, but not for Zebra. Such a discrepancy inherently degrades the effectiveness of zero-shot learning of Pig class. To our knowledge, this problem has neither been identified nor addressed in the zero-shot learning literature.

1.2.3 Prototype Sparsity Problem

This problem refers to the fact that for each target class, we only have a single prototype which is insufficient to fully represent what that class looks like. As shown in Figs. 1.2(b), there often exists large intra-class variations and inter-class similarities. Consequently, even if the single prototype is centered among its class members in the semantic representation space, existing zero-shot learning classifiers still struggle to assign the correct class labels to these highly overlapped data points – one prototype per class simply is not enough to represent the intra-class variability. This problem has never been explicitly identified although a partial solution exists [RES13].

1.2.4 Inability to Combine Multiple Semantic Representations

In addition to these inherent problems, conventional approaches to zero-shot learning are also limited in exploiting multiple intermediate semantic spaces/views, each of which may contain complementary information – they are useful in distinguishing different classes in different ways. In particular, while both visual attributes [LNH09, LNH13, FEHF09, LKS11, FHXG13] and linguistic semantic representations such as word vectors [MCCD13, FCS$^+$13, SGS$^+$13] have been independently exploited successfully, it remains unattempted and not straightforward to exploit synergistically multiple semantic 'views'. This is because they are often of very different dimensions and types and each suffers from different domain shift effects discussed above. This exploitation has to be transductive for zero-shot learning as only unlabelled data is available for the target classes and the labelled auxiliary data cannot be used directly due to the projection domain shift problem.

The problems discussed in Section 1.2.2, Section 1.2.3 and Section 1.2.4 are intertwined with each other. For example, in the widely used DAP model shown in Figure 1.3, different components of the model are affected by different problems and their negative effects aggregate and degrade the performance of zero-shot learning.

1.2.5 Noisy Annotations of Relative Attributes

With the advance of the Internet technology, some crowdsourcing tools are employed to collect a large volume of annotated attributes. Most user-defined attributes are binary, i.e. indicating the presence/absence of certain properties in image or video instances/classes. This limits the expressive power of attributes. For example, for some properties, there is not a clear case of

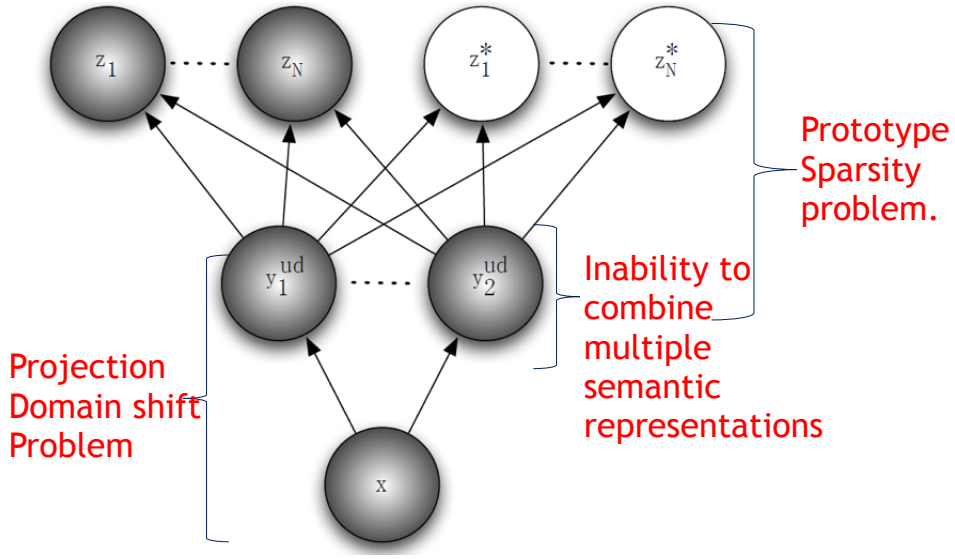


Figure 1.3: Projection domain shift problem, prototype sparsity problem and the inability to combine multiple semantic representations are tightly related and take negative effects in DAP model [LNH09, LNH13].

having them or not; it is more about how much the entity has such properties, i.e. these properties are continuous in nature. In other words, these binary attributes fail to capture more general semantic information, e.g., describing the relative relationship of any two instances.

To tackle this problem, relative attributes are studied recently. *Relative attributes* [PG11b] indicate the strength of the presence of certain visual properties (e.g. smiling, age for faces, and naturalness for scene). Relative attributes are learned as a richer representation corresponding to the strength of visual properties, and used in a number of tasks including visual recognition with sparse data, interactive image search, and semi-supervised or active learning of visual categories.

In existing work [KPG12, GGR+13, JYF+13], crowdsourced pairwise comparisons of relative attributes are collected. Crowdsourcing tools such as Amazon Mechanical Turk (AMT) are used to collect large-scale pairwise comparisons of relative attributes by asking participants to compare two instances for one particular attribute. The annotation task is to select between a pair of images or videos which one has more attributes. This is considered to be a much easier task and results in more reliable annotations than traditional five-star annotations[4]. This labelling pro-

[4]The five-star annotation system is widely studied in Multimedia and Statistical communities [XXHY13, CWCL09a, WCCL13, Arr63]. It requires the annotators to directly assign an absolute attribute value scaling from 1 to 5 (meaning weakest to strongest attribute) for each image/video instance. It has been taken as an extremely hard task [WCCL13, CWCL09a], since different people may hold very different understanding and judgement for the scale values.

cess is more economic and scalable than the conventional laboratory-based pairwise annotations.

Nevertheless, unlike the well-controlled and well-qualified laboratory annotators, there are many uncertainties in crowdsourcing scenarios including both diverse Internet annotators and the un-controlled crowdsourcing labelling process itself. These uncertainties introduce much higher ratios of noisy annotations than in laboratory annotation. These noisy annotations of relative attributes also suffer from two aspects,

Sparsity the number of pairwise comparisons required is much bigger than that of directly annotated relative attribute values due to n instances defining a $\mathcal{O}(n^2)$ pairwise space; even with crowdsourcing tools, the annotations will still be sparse, i.e. not all pairs are compared and each pair is only compared a few times.

Outliers it is well known that crowdsourced data is greatly affected by noise and outliers [CB13, WHG11, LHK13] which are caused by many different factors. As illustrated in Figure1.4(a), different annotators will give contradictory preference between Monkey King and Cookie monster due to different cultural and psychological factors. For example, if one annotator likes the story of "*Journal to West*", he/she would prefer Monkey King[5]. Ambiguous comparisons or malicious/lazy annotators are also reasons for noise and outliers. For example, the ambiguous smiling/crying left face in Figure1.4 (b) may bring noise/outlier comparisons to crowdsourced annotations.

Recent studies [GGR+13, KPG12, JYF+13] employ crowdsourcing tools to collect pairwise comparisons – relying on majority voting to prune the annotation outliers/errors. However, it is only a greedy algorithm and the performance of outlier detection by majority voting suffers from the sparsity problem.

1.3 Our Approach

In order to solve the limitations of user-defined attributes, we propose learning latent attributes by a novel generative topic model. To tackle the projection domain shift problem, prototype sparsity problem, and inability to combine multiple semantic representations, we learn a transductive multi-view embedding and achieve recognition by multi-view hypergraph label propagation. For noisy annotations of relative attributes, we propose an algorithm capable of robustly learning relative attributes.

[5]This can be psychologically explained as halo effect.

Figure 1.4: Pairwise comparisons of interesting and smiling.

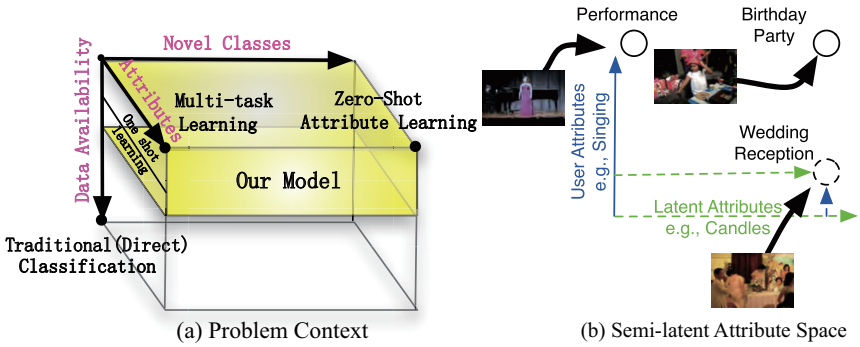


Figure 1.5: Problem context and semi-latent attribute space: (a) Learning a semi-latent attribute space is applicable to various problem domains. (b) Representing data in terms of a semi-latent attribute space partially defined by the user (solid axes), and partially learned by the model (dashed axes). A novel class (dashed circle) may be defined in terms of both user-defined and latent attributes.

1.3.1 Learning Latent Attributes

In order to break the limitations of user-defined attributes, we propose to *learn latent attributes*, which enables the task of attribute learning for understanding complex multimedia data with sparse and incomplete labels. In particular, our work can deal with the complex consumer videos of social group activities, which are challenging and topical examples for learning latent attributes because of their multi-modal content and complex and unstructured nature relative to the density of user-defined attribute annotations. We thus investigate the learning of latent attributes for content-based understanding, which aims to model and predict classes and tags relevant to objects, sounds and events – anything likely to be used by humans to describe or search for media.

One strategy to address the sparsity of user-defined attributes is via the exploitation of at-

tributes between different classes. As mentioned before, attributes focus on *describing* an instance (e.g., has legs) rather than *naming* it (e.g., is a dog), and they provide a semantically meaningful bridge between raw data and higher level classes. The concept of attributes can be traced back to the early work of intrinsic images [BT78], but attribute learning has been popularized recently as a powerful approach for image and video understanding with sparse training examples [LNH09, FEHF09, FEH10, PG11a, PHPM09]. Most such previous work have looked at attributes as a solution to *sparseness* of user-defined attributes, but do focus on constrained domains and single modalities, avoiding the bigger issues in intrinsic *incompleteness* and *ambiguity*.

To address these challenges, we introduce a new attribute learning framework (Figure 1.5) which learns a unified *semi-latent attribute space* (Figure 1.5(b)). *Latent attributes* represent all shared aspects of the data which are not explicitly included in users' sparse and incomplete annotations of user-defined attributes. These are complementary to user-defined attributes, and discovered automatically by a model through joint learning of the semi-latent attribute space.

This learned space provides a mechanism for *semantic feature reduction* [PHPM09] from the raw data in multiple modalities to a unified lower dimensional semantic attribute space (Figure 1.5(b)). The semi-latent space bridges the semantic gap with reduced dependency on the completeness of the attribute ontology and accuracy of the training attribute labels. Figure 1.5(a) highlights this property by putting our work in the context of various standard problems. Our semi-latent attribute space consists of three types of attributes: user-defined (UD) attributes from any prior concept ontology; latent class-conditional (CC) attributes [HLGX11] which are discriminative for known classes; and latent generalized free (GF) attributes [HGX11a] which represent shared aspects not in the attribute ontology. Jointly learning this unified space is important to ensure that latent CC and GF attributes represent un-modeled aspects of the data rather than merely rediscovering user-defined attributes.

To learn the semi-latent attribute space, we propose a multi-modal latent attribute topic model (M2LATM), building on probabilistic topic models [BNJ03, HLGX11]. M2LATM jointly learns user-defined and latent attributes, providing an intuitive mechanism for bridging the semantic gap and modeling sparse, incomplete and ambiguous labels. To learn the three types of attributes, the model learns three corresponding sets of topics with different constraints. UD topics are constrained to one to one correspondence with attributes from the ontology. Latent CC topics are

constrained to match the class label while latent GF topics are unconstrained. Multi-task classification, N-shot learning and zero-shot learning are performed in the learned semantic attribute space. To make the learning and inference scalable, we exploit equivalence classes for scalability by expressing our topic model in a "vocabulary" rather than "word" domain.

1.3.2 Transductive Multi-view Embedding

In order to solve the projection domain shift problem, prototype sparsity problem and the inability to combine multiple semantic representation together, we propose a *Transductive Multi-view Embedding framework* which has two major components: learning a transductive multi-view embedding and recognition by Multi-view Hypergraph Label Propagation (TMV-HLP).

Learning a transductive multi-view embedding The first component solves the projection domain shift problem inherent to previous attribute learning pipelines and exploits multiple semantic representations; each of which may contain complementary information. By using such complementary information, such a transductive framework can solve the prototype sparsity problem. Under our framework, each unlabelled instance from the target classes is represented by multiple views: its low-level feature view and its (biased) projections in multiple semantic spaces (visual attribute space, word space and so on). We introduce a multi-view semantic space alignment process to correlate different semantic views and the low-level feature view by projecting them onto a latent embedding space learned using multi-view Canonical Correlation Analysis (CCA) [GKIL13]. The objective of learning this new embedding space is to transductively (using the unlabelled target data) align the semantic views with each other, and with the low-level feature view to rectify the projection domain shift and exploit their complementarity. Even with the proposed transductive multi-view embedding framework, the prototype sparsity problem remains – instead of one prototype per class, a handful are now available depending on how many views are embedded, which are still sparse. Our solution to this problem is to explore the manifold structure of the data distributions of different views projected onto the same embedding space via label propagation on a graph. Thus, to solve it, we further present TMV-HLP in the embedding space.

Transductive multi-view hypergraph label propagation (TMV-HLP) The core of our TMV-HLP algorithm is a new *distributed representation* of graph structure termed a heteroge-

neous hypergraph – instead of constructing hypergraphs independently in different views (i.e. homogeneous hypergraphs), data points in different views are combined to compute multi-view heterogeneous hypergraphs. This allows us to exploit the complementarity of different semantic and low-level feature views, as well as the manifold structure of the target data to compensate for the impoverished supervision available in the form of the sparse prototypes. Zero-shot learning is then performed by semi-supervised label propagation from the prototypes to the target data points within and across the graphs.

By combining our transductive embedding framework and the TMV-HLP zero-shot recognition algorithm, our approach seamlessly generalises when none (zero-shot), or few (N-shot) samples of the target classes are available. Uniquely it can also synergistically exploit zero + N-shot (i.e., both prototypes and labelled samples) learning. Furthermore, the proposed method enables a number of novel cross-view annotation tasks including zero-shot class description and zero attribute learning.

1.3.3 Robust Learning of Relative Attributes

In order to learn and predict the image/video relative attributes from their low-level features and crowdsourced pairwise annotations, we propose the approach of *robust learning of relative attributes*. We show that the proposed approach is a principled way of identifying annotation outliers by formulating the whole task as a unified robust learning to rank problem which thus jointly tackles both the outlier detection and relative prediction.

We propose a novel approach for predicting relative attributes from sparse and noisy pairwise comparison data. In previous work, majority voting [GGR$^+$13, JYF$^+$13] or Huber-LASSO [XXHY13, FTS12] are used for outlier detection of pairwise comparisons and followed by regression [GGR$^+$13] or learning to rank [JYF$^+$13]. However, majority voting is a local and greedy algorithm for outlier detection. It needs lots of pairwise comparisons and cannot guarantee the performance of outlier detection. Different from existing approaches, we formulate a unified robust learning to rank framework to jointly solve both the outlier detection and the prediction of relative attributes. Critically, instead of detecting outliers locally and independently at each pair by majority voting, our outlier detection method operates globally, integrating all local pairwise comparisons together to minimise a cost that corresponds to global inconsistency of ranking order. This enables us to identify outliers that receive majority votes and yet cause large global ranking inconsistency

and thus should be removed. Furthermore, as a global method, only one comparison per pair is required; therefore significantly reducing the data sparsity problem compared to the conventional majority voting approach.

1.4 Contributions

The contributions of this thesis towards attribute learning for image and video understanding are as follows,

1. We address the key limitation of user-defined attributes, i.e., attribute learning from sparse, incomplete and ambiguous annotations of user-defined attributes. A semi-latent attribute space is introduced and enables the use of as much or as little prior knowledge as available from both user-defined and the two types of automatically discovered latent attributes. We formulate a computationally tractable solution of this strategy via a novel and scalable topic model. We also show how latent attributes computed by our framework can be utilised to tackle a wide variety of learning tasks in the context of multimedia content understanding including multi-task, label-noise, N-shot and surprisingly zero-shot learning.

2. For the first time we attempt to investigate and provide a solution to the projection domain shift problem in zero-shot learning. A transductive multi-view embedding space is learned that not only rectifies the projection shift, but also exploits the complementarity of multiple semantic representations of visual data.

3. The prototype sparsity problem can also be tackled in our transductive multi-view embedding framework. A novel transductive multi-view heterogeneous hypergraph label propagation algorithm (TMV-HLP) is developed to improve both zero-shot and N-shot learning tasks in the embedding space and overcome the prototype sparisity problem.

4. Our transductive multi-view embedding space enables the novel task: zero-shot annotation. It includes zero-shot class description (inferring the semantic attribute description of a novel class) and zero attribute learning (inferring the name of a novel class given a set of attributes).

5. We propose a learning framework to robustly predict relative attributes. Such a framework can learn relative attributes from noisy and sparse pairwise comparison data. For the first time, the problems of detecting outliers and estimating the ranking score are solved

jointly in our unified framework. Most importantly, from both theoretical and experimental aspects, we demonstrate that our method is superior to existing majority voting based methods as well as statistical ranking based (e.g. huber-LASSO) methods.

6. The first unstructured multi-modal social activity attribute (USAA) dataset is contributed. We manually annotated the groundtruth attributes for 8 semantic class videos, which are birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception. We also define 69 multi-modal attributes which can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement. We tried our best to exhaustively define every conceivable attribute for this dataset, to make a benchmark for unstructured social video classification and annotation. Such exhaustive annotations give the freedom to hold out various subsets and learn on the others in order to quantify the effect of annotation density and biases on a given algorithm. Thus this dataset is of great value to computer vision community. We will further discuss the USAA dataset in Chapter 2.1.6.

1.5 Outline

This thesis is organised into the following chapters,

Chapter 2 presents the literature review on various existing attribute learning methods in computer vision, summarising the datasets used in the thesis, and related work in machine learning used in this thesis.

Chapter 3 provides detailed explanations of the learning latent-attribute framework. It shows that our framework can jointly learn multi-modal user-defined and latent attributes that enable automatic video classification and annotation of unstructured group social activity in videos.

Chapter 4 explains the transductive multi-view embedding frameworks. Transductive multi-view embedding zero-shot learning has two major components, i.e., learning a transductive multi-view embedding and recognition by Multi-view Hypergraph Label Propagation (TMV-HLP). It can be used to solve the projection domain shift problems, prototype sparsity problems and the inability to combine multiple semantic views.

Chapter 5 presents novel framework for robust learning of relative attributes from noisy crowd-sourced data. We show that the presented framework can effectively identify outliers for robust relative attribute learning with extremely noisy and sparse annotations.

Chapter 6 provides conclusions and suggests a number of areas to be pursued in the future.

Chapter 2

Literature Review

In this chapter, we review some topics related to *attribute learning on image and video understanding*. Firstly, we briefly look at previous attribute learning work in computer vision community in Section 2.1; Secondly, Section 2.2 talks about some other semantic representations beyond attributes. Thirdly, the machine learning approaches which are related to the algorithms in this thesis are reviewed in Section 2.3. Finally, we summarize the whole chapter in Section 2.4.

2.1 Attribute Learning in Computer Vision

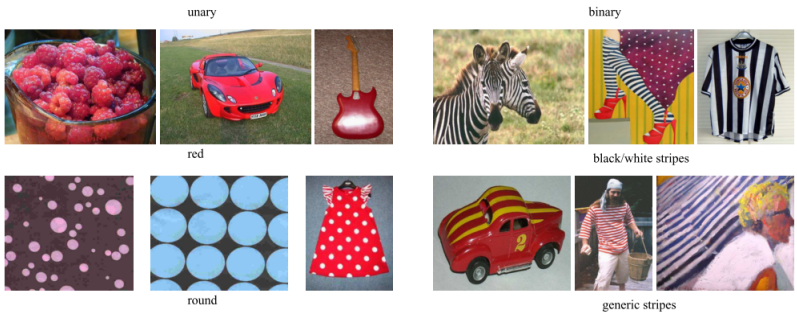


Figure 2.1: Examples of different kinds of attributes. The 'unary' attributes indicate some simple attributes, whose characteristic properties are captured by individual image segments (appearance for red, shape for round). In contrast, the 'binary' attributes are more complex attributes, whose basic element is a pair of segments (e.g. black/white stripes). Images from [FZ07].

The word attribute (e.g., has wings) prefers to the intrinsic characteristic that embody an in-

stance or a class (e.g., bird) (Fu *et al.* [FHXG12]) and a human has the ability to decide whether such a characteristic is present or not for a certain object (Lampert *et al.* [LNH13]). Thus attributes answer the question of describing a class or instance in contrast to the typical (classification) question of naming an instance. The attribute description of an instance or category is useful as a semantically meaningful intermediate representation to bridge the gap between low level features and high level class concepts (Palatucci *et al.* [PHPM09]).

2.1.1 Attribute Learning Models

We will briefly review several of the most commonly used attribute learning models in this section. Generally speaking, a key advantage of attribute learning models is their use to provide an intuitive mechanism for multi-task (Salakhutdinov *et al.* [STT11]) and transfer learning (Hwang *et al.* [HSG11]): enabling learning with few or zero instances of each class via sharing attributes – zero-shot/N-shot learning. Particularly, the challenge of zero-shot recognition (as illustrated in Figure 2.2) is to recognize unseen visual object categories without any training exemplars of the unseen class. This requires the transfer of knowledge of additional semantic information from auxiliary classes with example images to unseen target classes.

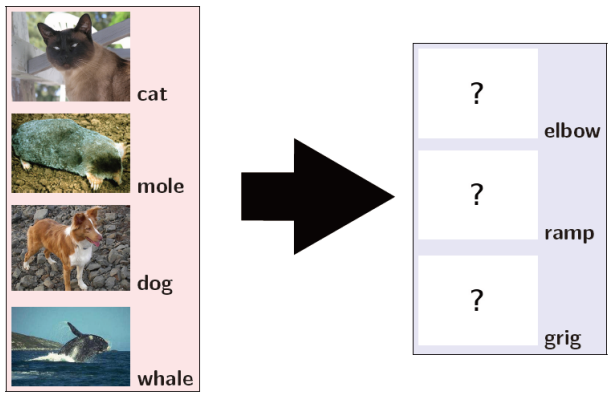


Figure 2.2: To recognise novel classes, zero-shot learning transfers knowledge from classes with examples to novel classes. Images from Dr. Christoph Lampert's slides for [LNH09].

Attribute learning models have been explored for images and to a lesser extent video (Liu *et al.*[LKS11] and Fu et al. [FHXG12, FHXG13] as well as Chapter 3). Applications include modeling the properties of human actions (Liu *et al.* [LKS11]), animals (Lampert *et al.* [LNH09, LNH13]), faces (Kumar *et al.* [KBBN09]), scenes (Hwang *et al.* [HSG11]), and objects (Farhadi *et al.* [FEH10, FEHF09]). Most of these studies assumed that an exhaustive space of attributes

Figure 2.3: The high-level attributes allows the transfer of knowledge between object categories [LNH09]: the visual appearance of attribute is independently learned from training examples and across different categories; the the object class without any training images can be detected based on which attribute description a test image fits best. Images from [LNH09].

has been manually specified.

Generative models for visual attributes The earliest work on attributes in Ferrari *et al.* [FZ07] studied some elementary properties such as colour or geometric pattern. From human annotations, Ferrari and Zisserman in [FZ07] proposed a generative model for learning simple color and texture attributes. Specifically, we use model $\mathcal{M}$ to explain a whole image $\mathcal{I}$. And the image $\mathcal{I}$ is further represented by a set of segments $\{s\}$. A latent variable f is defined to be associated with a foreground $(f = 1)$ or background $(f = 0)$ segment. All f for all segments of $\mathcal{I}$ are grouped into a vector $\mathbf{F}$. So the likelihood of the image is

$$p(\mathcal{I}|\mathcal{M};\mathbf{F},a) = \prod_{s\in I} p(s|\mathcal{M};f,a)^{N_s} \tag{2.1}$$

where N_s is the number of pixel the image contains. Different types of attributes will configure distinctive probability formulations which are specified by parameter $\mathcal{M}$. For example, as illustrated in Figure 2.1, the attribute can be either viewed as an unary (e.g. red colour and round texture), or a binary (e.g., black/white stripes).

Some later work (Parikh et a.[PG11b], Kovashka *et al.* [KPG12] and Berg *et al.* [BBS10]) extended the unary/binary attributes to compoundable attributes, which makes them extremely

useful for information retrieval (e.g., complex queries such as "Asian women with short hair, big eyes and high cheekbones") and identification (e.g., finding an actor whose name you forgot, or an image that you have misplaced in a large collection).

The generative models are formulated as the Bayesian formulation and enable the easy integration of prior knowledge of each type of attribute to compensate for limited supervision in image and video understanding. The framework proposed in Chapter 3 belongs to the category of generative models. In that framework, we have different prior knowledge for user-defined and data-driven attributes.

IAP and DAP models Lampert *et al.* [LNH09, LNH13] studied the problem of object recognition of categories for which no training examples are available. To solve such a problem, attribute-based classification is introduced to perform object detection based on an intermediate level semantic attribute representations. As illustrated in Figure 2.3, such attributes transcends the specific learning tasks and pre-learned independently across different categories and thus allowing transferring knowledge. Specifically, for zero-shot learning tasks, they proposed two probabilistic frameworks, i.e., Direct Attribute Prediction (DAP) in Figure2.4(b) and Indirect Attribute Prediction (IAP) in Figure2.4(c), these models can integrate human knowledge in the recognition process of unseen classes by using category-level class-attribute associations.

- *DAP model* Assume the relation between known classes $y_i, ..., y_k$, unseen classes $z_1, ..., z_L$ and descriptive attributes $a_1, ..., a_M$ is given by the matrix of binary associations values a_m^y and a_m^z. Such a matrix encodes the status of one attribute regarding one given class. Extra knowledge is applied to define such an association matrix, for instance, by human experts (Lampert *et al.* [LNH09, LNH13]), by concept ontology (Fu *et al.* [FHXG13]), and by semantic relatedness measured between class and attribute concepts (Rohrbach *et al.* [RSS12]). In the training stage, the attribute classifiers are trained by the attribute annotations of known classes $y_i, ..., y_k$. At the test stage, the posterior probability $p(a_m|x)$ can be inferred for an individual attribute a_m in an image x. To predict the class label of object class z,

$$p(z|x) = \Sigma_{a\in\{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m|x)^{a_m^z} \tag{2.2}$$

- *IAP model* The DAP model directly learns attribute classifiers from the known classes, while the IAP model builds attribute classifiers by combining the probabilities of all associated known classes. It is also introduced as direct similarity-based model in Rohrbach

et al. [RSS12]. In the training step, we can learn the probabilistic multi-class classifier to estimate $p(y_k|x)$ for all training classes $y_i, ..., y_k$. Once $p(a|x)$ is estimated, we use it as the same way as in for DAP in zero-shot learning classification problems. In the testing step, we predict,

$$p(a_m|x) = \Sigma_{k=1}^{K} p(a_m|y_k)p(y_k|x) \tag{2.3}$$

PST model Rohrbach *et al.* [RES13] explored the manifold structure of the instances in the novel classes to help attribute-based transfer learning for zero-shot and N-shot learning. Thus they proposed a graph-based semi-supervised learning algorithm – PST model. Specifically, they constructed a k-NN graph by using the low-level features of testing data. The distance of any two data pairs (x_i, x_j) is

$$d(x_i, x_j) = \sum_{d=1}^{D} \left| x_{i,d} - x_{j,d} \right|$$

where D is the dimensionality of the low-level feature space. Once the k-NN graph is computed, they replace the original distance with the semantic distance of attribute vectors by using

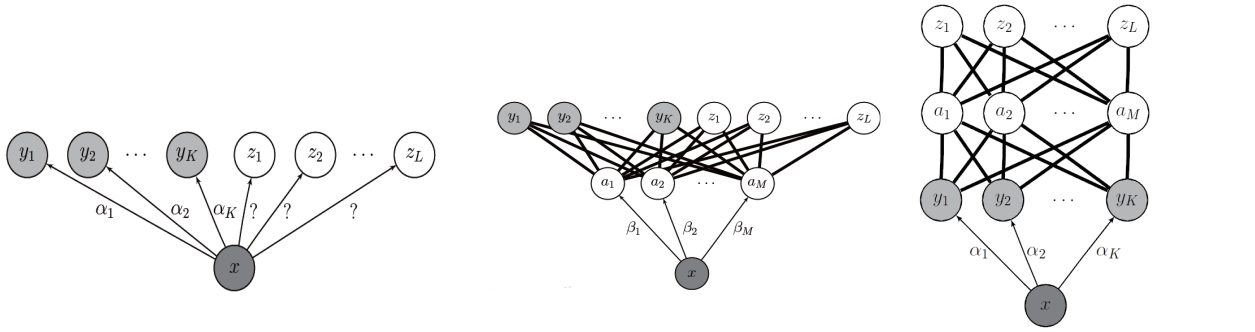
$$d(x_i, x_j) = \sum_{m=1}^{M} |p(a_m|x_i) - p(a_m|x_j)|$$

where $M \ll D$ and the similarity of the whole graph is measured by the RBF kernel. The label set is initialized by the nearest neighbourhood distance of each testing instance to the prototypes of each novel class.

2.1.2 Binary Vs. Relative Attributes

The attributes discussed above are 'binary'[1] and they may be sufficient to indicate some properties (e.g. spotted) or annotations (e.g. has a head) of one image or object. Farhadi *et al.* [FEHF09] learned a richer set of attributes including parts, shape, materials and etc. Another commonly used methodology (e.g. Liu *et al.* [LKS11] in human action recognition, and Wang *et al.* [WZ11] in attribute and object-based model) is to take the attribute labels as latent variables on the training dataset in the form of a structured latent SVM model, and the objective is to minimize object prediction loss. In contrast, relative information in the form of relative attributes can

[1]Note that 'binary' indicates using a single value to represent the strength of the attribute on one instance/class. Nevertheless, such a value is not necessarily binary.



(a) Flat multi-class classification (b) Direct attribute prediction (DAP) (C) Indirect attribute prediction (IAP)

Figure 2.4: Graphical representation of three different models: Flat, DAP and IAP. Dark gray nodes (images x) are always observed; light gray nodes (auxiliary classes: y_1,y_2, $\cdots$, y_K) are observed only during training. And white nodes (novel target classes z_1, z_2,$\cdots$, z_L) are not observed but must be inferred. DAP and IAP indicate two different ways for zero-shot learning, while Flat model cannot generalise to detect novel target classes of no training examples. Images from [LNH09].

be used as a more informative way to express richer semantic meaning and thus better represent visual information.

Relative attributes (Parikh *et al.* [PG11b]) were recently proposed to learn a ranking function to predict the relative semantic strength of image attributes. Specifically, annotators give pairwise comparisons on images and a ranking function is then learned to estimate relative attribute values for unseen images as ranking scores. These relative attributes are learned as a richer representations corresponding to the strength of visual properties, and used in a number of tasks including visual recognition with sparse data, interactive image search (Kovashka *et al.* [KPG12]), and semi-supervised (Shrivastava *et al.* [SSG12]) or active learning (Biswas *et al.* [BP13, PP12]) of visual categories. Kovashka *et al.* [KPG12] proposed a novel model of feedback for image search where users can interactively adjust the properties of exemplar images by using relative attributes in order to best match his/her ideal queries. This searching process is illustrated in Figure 2.6.

In a broader sense, many other tasks of estimating continuous values representing visual properties in image/video are also examples of relative attribute learning, e.g., image/video interestingness [GGR$^+$13, JYF$^+$13], memorability [IPTO11, IXTO11], aesthetic [DOB11], and human-face age estimation [FGH10, CGXL13]. As one special case of relative attribute, image interestingness is studied by Gygli *et al.* [GGR$^+$13], which showed that three cues contribute

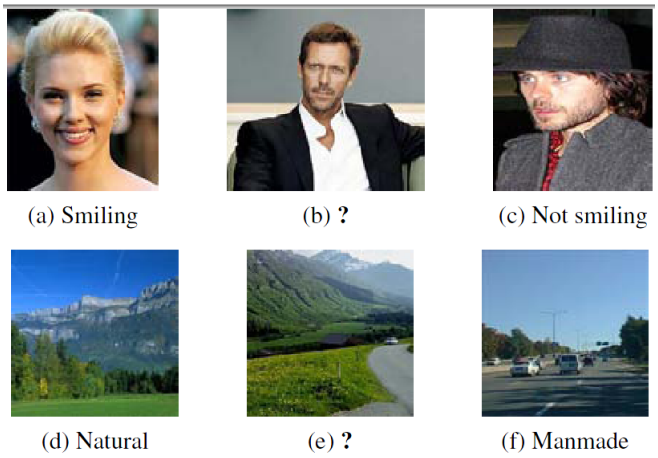


Figure 2.5: Relative attribute [PG11b] enables semantic richer textual descriptions than binary attribute. For example, (b) is smiling more than (c) but less than (c); (e) is less natural than (d) but more than (f).

the most to interestingness: aesthetics, unusualness/novelty and general preferences, the last of which refers to the fact that people in general find certain types of scenes more interesting than others, for example outdoor-natural vs. indoor-manmade. As the first work on predicting video interestingness (illustrated in Figure 2.12), Jiang *et al.* [JYF+13] evaluated the different features for video interestingness prediction from crowdsourced pairwise comparisons.

2.1.3 User-defined Vs. Data-driven Attributes

The attributes are usually defined by extra-knowledge of either expert users or concept ontology. To better augment such user-defined attributes, Parikh *et al.* [PG11a] proposed a novel approach to actively argument the vocabulary of attributes to both help resolve intra-class confusions of new attributes and coordinate the "nameability" and "discriminativeness" of candidate attributes. However, such user-defined attributes are far from enough to model the complex visual data. The definition process can still be either inefficient (costing substantial effort of user experts) and/or insufficient (descriptive properties may not be discriminative). To tackle such problems, it is necessary to automatically discover more discriminative intermediate representations from visual data, i.e. data-driven attributes.

Data-driven attributes have only been explored in a few previous works. Liu *et al.* [LKS11] employed an information theoretic approach to infer the data-driven attributes from training examples by building their framework on a latent SVM formulation. They directly extended the

Figure 2.6: Whittle search [KPG12]: Example iterative search results with relative attribute feedback.

attribute concepts in images to comparable "action attributes" in order to better recognize human actions. Attributes are used to represent human actions from videos and enable the construction of more descriptive models for human action recognition. They augmented user-defined attributes by data-driven attributes, similar to latent class-conditional (CC) attributes learned in Chapter 3, to better differentiate existing classes. However, our more nuanced distinction between CC and latent generalized free (GF) latent attributes better helps differentiate both existing classes and novel classes: class-conditional attributes are limited to those which differentiate existing classes; without this constraint, GF attributes provide an additional cue to help differentiate novel classes. Farhadi *et al.* [FEHF09] also learned user-defined and CC attributes separately. This means that the learned CC attributes are not necessarily complementary to the user-defined ones (i.e., they may be redundant).

Another limitation of previous data-driven attribute work [FEHF09, LKS11] is that their data-driven attributs can not be directly used in zero-shot learning. This limits the efficacy of learning the data-driven attributes. In Chapter 3, we uniquely showed how to use latent attributes in zero-shot learning.

Despite such previous efforts, an exhaustive space of attributes is unlikely to be available, due to the expense of ontology creation, and semantically obvious attributes for humans do not necessarily correspond to the space of detectable and discriminative attributes. One method of collecting labels for large scale problems is to use Amazon Mechanical Turk (AMT) [SF08]. However, even with excellent quality assurance, the results collected still exhibit strong label

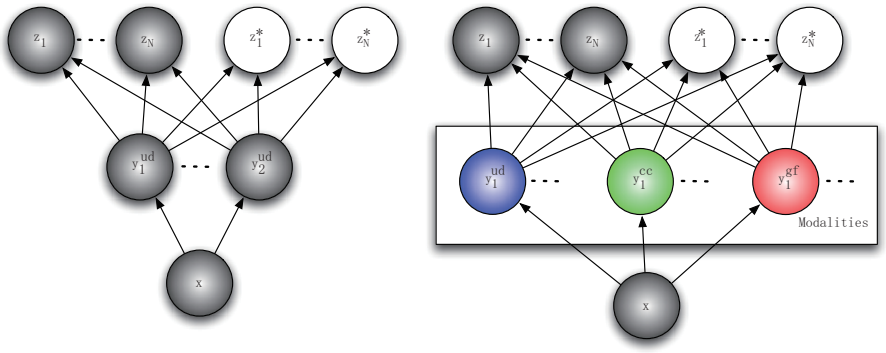


Figure 2.7: Schematic of conventional (left) DAP [LNH09] versus (right) M2LATM [FHXG13]. Shading indicates different types of constraints placed on the variables.

noise. Thus label-noise [TYH+09] is a serious issue in learning from either AMT, or existing social meta-data. More subtly, even with an exhaustive ontology, only a subset of concepts from the ontology are likely to have sufficient annotated training examples, so the portion of the ontology which is effectively usable for learning, may be much smaller. To solve this problem, in Chapter 3 and [FHXG12, FHXG13] we introduce the concept of a semi-latent attribute space, expressing user-defined and latent attributes in a unified framework, and proposed a novel scalable probabilistic topic model for learning multi-modal semi-latent attributes, which dramatically reduces requirements for an exhaustive and accurate attribute ontology and expensive annotation effort. Figure 2.7 contrasts Direct Attribute Prediction (DAP [LNH09]) with our multi-modal latent attribute topic model (M2LATM) [FHXG13]. The shading indicates the types of constraints placed on the nodes, with the dark nodes being fully observed, and the colored nodes in M2LATM having user-defined, CC and GF type constraints.

2.1.4 Image Vs. Video Attributes

As discussed, most previous work in Section 2.1.1, Section 2.1.2 and Section 2.1.3 focus on image attributes. However, most previous attribute work focused on image understanding, few are about video attributes on video understanding.

Video attributes are recently studied and aim to indicate a wide range of topics such as those related to objects (e.g., animal), indoor/outdoor scenes (e.g., meeting, snow), events (e.g., wedding ceremony), and so on. Figure 2.8 gives one example of video attributes. Video attributes thus share many similarities with the video concept detection in Multimedia community, for example, recent studies address video concept detection (Snoek *et al.* [SHH+07] and Hauptmann

Figure 2.8: The key frame of video shot can be represented by five attributes: "face," "people," "walking running," "people marching," and "flag". Images from [ZMWH07].

et al. [HYL+07]) (also known as tagging [HGX11a, TAP+10, YHWZ11], and multi-label classification (Qi *et al.*[QHR+07]), video annotation (Tang *et al.* [THW+09])).

Most video attributes refer to a video ontology. Depending on the ontology, the level of abstraction and models used, many annotation approaches can therefore be seen as addressing a sub-task of attribute-learning. Some annotation studies aim to automatically expand (e.g. Hauptmann *et al.*[HYL+07]) or enrich (Yang et al. [YHWZ11]) the set of tags queried in a given search. Nevertheless, the possible space of expanded/enriched tags is still constrained by a fixed ontology and may be very large (e.g., a vocabulary space of over 20,000 tags in Toderici *et al.* [TAP+10]).

2.1.5 Low-level Features

Since the performance of computer vision algorithms is heavily depending on the choice of data representation, we briefly summarise the low-level features used in this section. To detect visual attributes, the state-of-the-art attribute classifiers have to learn Supporter Vector Machine (SVM) classifiers from manually labelled images/videos which are represented by visual code features (Sande *et al.* [vdSGS10] and Jiang et al. [JYN10]). Such features can be global features, local features and recent deep features (Sermanet et al.[SEZ+14] and Donahue et al. [DJV+14]).

Global features (e.g. colour histogram, gist, edge, and wavelet) are statistics about the overall distribution of color, texture, or edge information. The global features are the most classical type of features and have been used in almost all earlier vision work. However, global features are

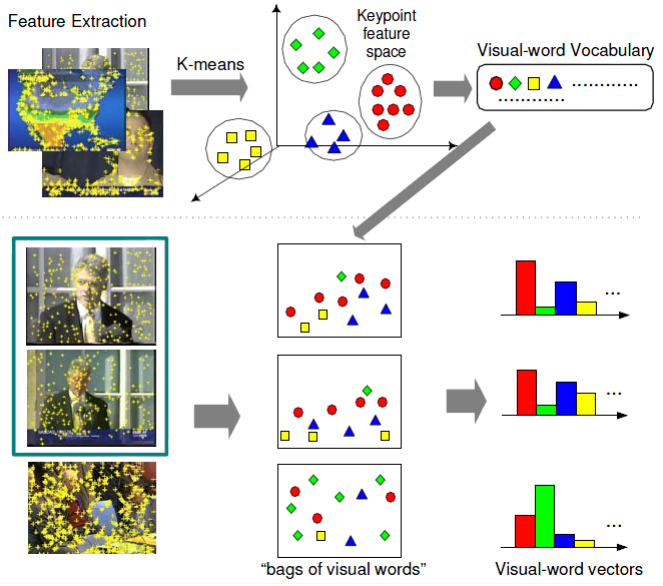


Figure 2.9: Image representation using bag-of-visual-words. Images from [JYN10].

generally limited due to the presence of huge within-class variation, pose and lighting changes in images and video shots.

To generate more robust low-level representations of image/video, features are extracted from the local structures of visual objects. Thus local features have been widely studied over the past decade. The most popular approach is the bag-of-visual-words (BoW) [SZ03]. As illustrated in Figure 2.9, a visual vocabulary is generated by grouping similar local interest points (e.g. SIFT [Low04], rgSIFT [vdSGS08], PHOG [BZM07b], SURF [BETG08], local self-similarity histograms [SI07]) into a large number of clusters; one cluster corresponding to one visual word. The image/video can thus be represented by a histogram of visual words which can be taken as features.

Recent developed OverFeat (Sermanet *et al.* [SEZ+14]) and DeCaf (Donahue *et al.* [DJV+14]) features are generated by training Convolutional Neural Networks on the large-scale ImageNet dataset [DDS+09] and achieve better or competitive results compared to the state-of-the-art on various dataset. A more complete review on deep features are beyond the scope of this thesis. We refer to Bengio *et al.* [BCV13] as a more thorough survey on this topic.

2.1.6 Attribute Learning Datasets

This section briefly summarises the dataset used in this thesis for attribute learning.

2.1.6.1 Animal with Attribute (AwA) dataset

AwA dataset is firstly proposed in [LNH09]. The 50 Osher-son/Kemp animal images are collected online. There are 30475 images with at least 92 examples of each class. Seven different feature types are provided: RGB color histograms, SIFT [Low04], rgSIFT [vdSGS08], PHOG [BZM07b], SURF [BETG08], local self-similarity histograms [SI07] and DeCaf [DJV+14]. The AwA dataset defines 50 classes of animals, and 85 associated attributes (such as furry, and has claws). Some examples are shown in Figure 2.3.

For the consistent evaluation of attribute-based object classification methods, the AwA dataset defined 10 test classes: *chimpanzee*, *giant panda*, *hippopotamus*, *humpback whale*, *leopard*, *pig*, *raccoon*, *rat*, *seal*. The 6180 images of those classes are taken as the test data, whereas the 24295 images of the remaining 40 classes can be used for training.

2.1.6.2 CUB-200-2011 dataset

CUB-200-2011 Wah *et al.* [WBW+11] contains 11788 images of 200 bird classes, as illustrated in Figure 2.10. This is a more challenging dataset than AwA – it is designed for fine-grained recognition and has more classes but fewer images. All images are annotated with bounding boxes, part locations, and attribute labels. Images and annotations were filtered by multiple users of Amazon Mechanical Turk. CUB-200-2011 is used as the benchmarks dataset for multi-class categorization and part localization. Each class is annotated with 312 binary attributes derived from the bird species ontology. We use 150 classes as auxiliary data, holding out 50 as target data, which is the same setting adopted in Akata *et al.* [APHS13].

2.1.6.3 Image interestingness dataset

The image interestingness dataset was first introduced in Isola *et al.* [IXTO11] for studying memorability, as illustrated in Figure 2.11. It was later re-annotated as an image interestingness dataset by Gygli *et al.*[GGR+13]. It consists of 2222 images, each represented as a 932 dimensional feature vector as in [GGR+13]. 16000 pairwise comparisons were collected by using AMT and are used as annotation.

2.1.6.4 Video interestingness dataset

The video interestingness dataset is the YouTube interestingness dataset introduced in Jiang *et al.* [JYF+13], which contains 14 different categories, each of which has 30 YouTube videos. $10 \sim 15$ annotators were asked to give complete interesting comparisons for all the videos in each category. So the original annotation is noisy but not sparse. We use a bag-of-words of

Figure 2.10: CUB-200-2011 dataset.

Figure 2.11: Image Interestingness dataset. Images from [IXTO11].

Scale Invariant Feature Transform (SIFT) and Mel-Frequency Cepstral Coefficient (MFCC) as the feature representation which are shown to be effective in Jiang *et al.*[JYF+13] for predicting video interestingness. An example of the video interestingness dataset is shown in Figure 2.12.

2.1.6.5 *Outdoor Scene Recognition (Scene) Dataset and Public Figure Face Database (PubFig)*

PubFig contains 772 images from 8 people and 11 attributes ('smiling', 'round face', etc.), and it averages 418 labelled pairs for each attribute from 241 training images. Some example images are shown in Figure 2.13. Scene (Oliva *et al.* [OT01]) consists of 2688 images from 8 categories and 6 attributes ('openness', 'natrual' etc.) and an average 426 labelled pairs for each attribute from 240 training images. Some examples are shown in Figure 2.14. Graphs constructed are thus extremely sparse. Pairwise attribute annotation was collected by AMT (Kovashka *et al.* [KPG12]). Each pair was labelled by 5 workers to average the comparisons by majority voting by Kovashka *et al.*[KPG12]. Gist [OT01] and colour histograms features are used for PubFig, and Gist alone for Scene. Each image also belongs to a class (celebrity or scene type).

2.1.6.6 *Unstructured Social Activity Attribute (USAA) dataset*

It is the first benchmark video attribute dataset for social activity video classification and annotation introduced by us in [FHXG12]. We manually annotated the groundtruth attributes for 8 semantic class videos of Columbia Consumer Video (CCV) dataset [JYC+11], and select 100

Figure 2.12: Video interestingness: Example frames of videos from the Flickr and YouTube datasets collected in [JYF+13]. For each dataset, the video on top is considered more interesting than the other one according to human judgements.

videos per-class for training and testing respectively. These classes were selected as the most complex social group activities. As shown in Figure 2.16, a wide variety of attributes have been annotated. By referring to the existing work on video ontology [ZMWH07, JYC+11], the 69 attributes can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement. We tried our best to exhaustively define every conceivable attribute for this dataset, to make a benchmark for unstructured social video classification and annotation. Real-world videos will not contain such extensive tagging. However, this exhaustive annotations give the freedom to hold out various subsets and learn on the others in order to quantify the effect of annotation density and biases on a given algorithm.

These eight classes are birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception (shown in Figure 2.15). Each class has a strict semantic definition in the CCV video ontology. Directly using the ground-truth attributes (average annotation density 11 attributes per video) as input to a SVM, the videos can be classified with 86.9% accuracy. This illustrates the challenge of this data: while the attributes are informative, there is sufficient intra-class variability in the attribute-space, that

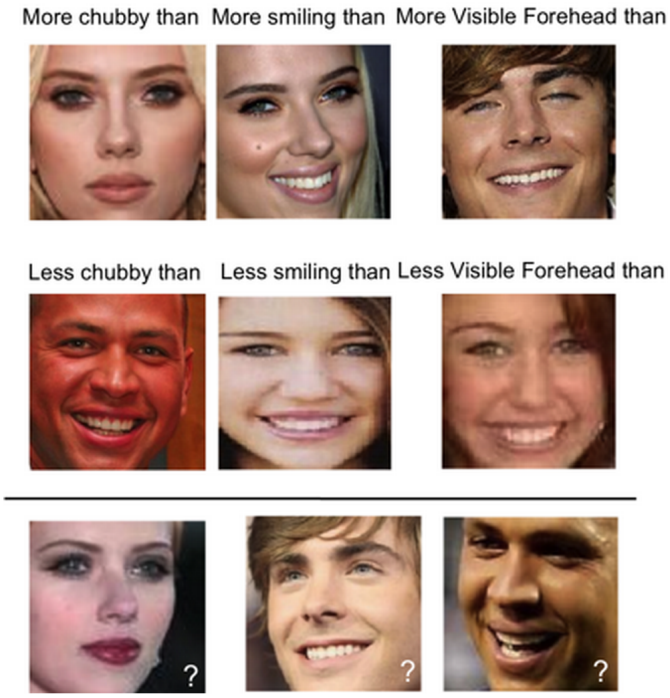


Figure 2.13: Some examples of PubFig dataset. Images from [PG11b].

Figure 2.14: Examples of OSR dataset.

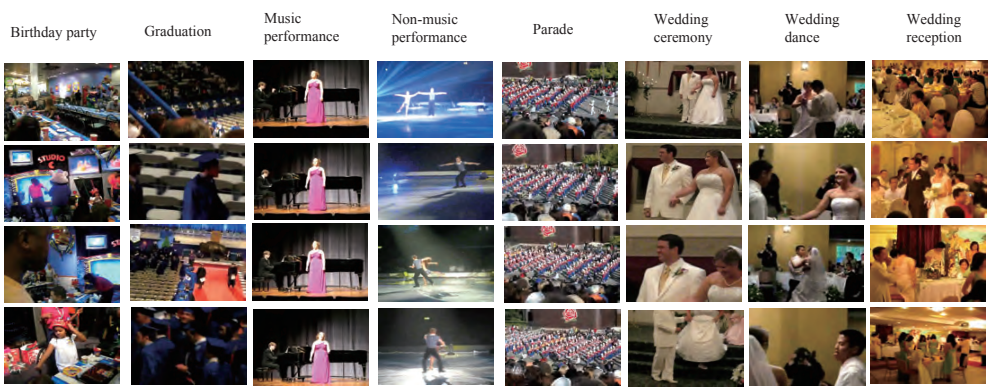


Figure 2.15: Example frames from the eight class unstructured social activity dataset.

Figure 2.16: Attribute examples in social activity attribute video dataset. Different types of attributes of both visual and audio modalities are shown in different colour.

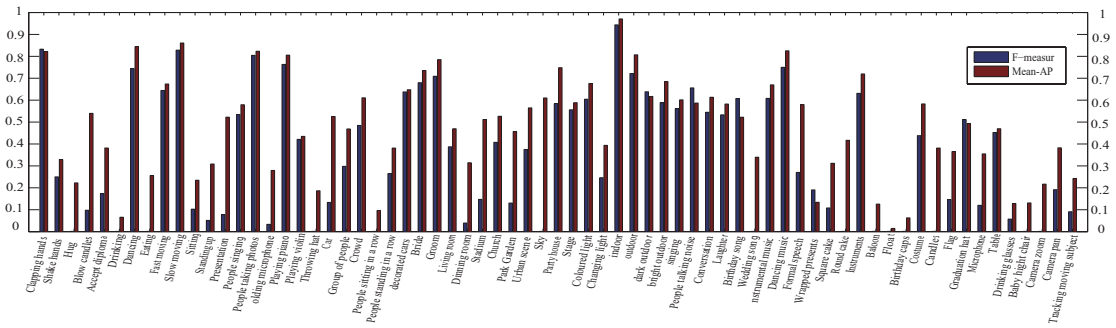


Figure 2.17: Attribute-classification accuracy using SVM on USAA dataset.

even perfect knowledge of the attributes in an instance is insufficient for perfect classification. The SIFT, STIP and MFCC features for all these videos are extracted according to [JYC+11], and included in the dataset. We report the baseline accuracy of SVM-attribute classifiers learned on the whole test set in Fig 2.17. Clearly some can be detected almost perfectly, and others cannot be detected given the available features.

2.2 Semantic Representations Beyond Attributes

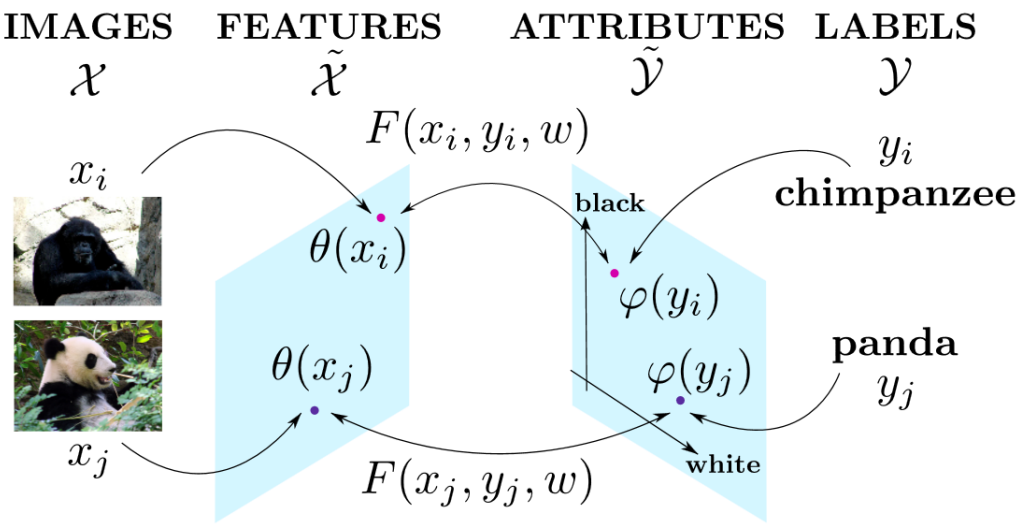


Figure 2.18: Akata *et al.* [APHS13] used attributes as side information for the label embedding and measured the "compatibility"' between the embedded inputs and outputs with a function F. Images from [APHS13].

There are many other semantic representations, e.g. semantic word vector and WordNet. The basic idea is to learn a semantic embedding function, i.e. a transformation, from the image (feature) space to a semantic space within which the unseen novel target classes can be defined and recognised by extra semantic knowledge. In the light of this idea, the attribute learning discussed above can be taken as one special case of such semantic embedding by mapping images to the semantic attribute space.

Larochelle *et al.* [LEB08] embedded handwritten character with a typed representation which

further helps recognise unseen classes. Specifically, once the embedding is learned from known classes, novel classes can be identified based on the similarity of prototype representations of classes and predicted presentations of the instances in the embedding space.

Linguistic semantic word space can also be employed for the semantic embedding for attribute learning. Socher *et al*. [SGS+13] learned a neural network model to embed each image with a 50-dimensional word vector in the semantic space which is trained by the unsupervised linguistic model [HSMN12] and using Wikipedia text. The images from either known or unknown classes could be mapped into such word vectors and classified by finding the closest prototypical linguistic word in such semantic space. Frome *et al*. [FCS+13] further scaled such ideas to recognise large-scale datasets. They proposed a deep visual-semantic embedding model to map images into a rich semantic embedding space for large-scale zero-shot recognition. Skip-gram model [BPV+92, MCCD13] was trained by a text corpus of 5.7 million documents (5.4 billionwords) from online encyclopedia (*wikipedia.org*) and used to construct such semantic embedding space. Different from the unsupervised linguistic model [HSMN12], such a skip-gram model is a re-current neural network model and models the syntactic and semantic regularities in language [23] which allows vector-oriented reasoning. Fu *et al.* [FYH+14b] showed that such a reasoning could be used to synthesize all different label combination prototypes in the semantic space and thus is crucial for multi-label zero-shot learning. For example, $Vec(\text{“}Moscow\text{”})$ should be much closer to $Vec(\text{“}Russia\text{”}) + Vec(\text{“}capital\text{”})$ than $Vec(\text{“}Russia\text{”})$ or $Vec(\text{“}capital\text{”})$ only in the semantic space.

Concept ontology is directly used as the semantic representation alternative to attributes. For example, WordNet [Mil95] is one of the most widely studied concept ontology. It is a large-scale semantic ontology built from a large lexical dataset of English. Especially, nouns, verbs and adjectives and adverbs are grouped into sets of cognitive synonyms (synsets) which indicates one distinct concept. Such an idea of semantic distance defined by the WordNet ontology is also used by Rohrbach *et al*. [RSS12, RSS+10] for transferring semantic information in zero-shot learning problems. They thoroughly evaluated many alternatives of semantic links between auxiliary and target classes by exploring linguistic bases such as WordNet, Wikipedia, Yahoo Web, Yahoo Image, and Flickr Image. Additionally, WordNet has been used for many vision problems. Fergus *et al*. [FBWT10] leveraged the WordNet ontology hierarchy to define semantic distance between any two categories for sharing labels in classification, as shown in Figure 2.19.

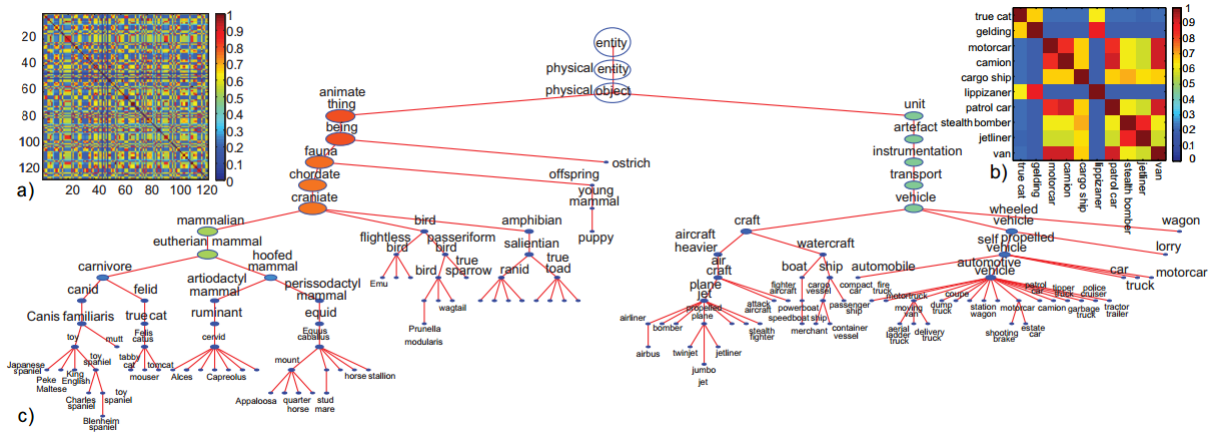


Figure 2.19: Examples of Wordnet sub-tree for a subset of 386 classes in [FBWT10]. The associated semantic affinity matrix is shown in (a), along with a closeup of 10 randomly chosen rows and columns in (b). Images from [FBWT10].

Semantic embedding by using attributes is another type of semantic representation different to semantic attributes. Akata *et al.* [APHS13] proposed to take attribute-based image classification as a label-embedding problem by minimising the compatibility function between an image and a label embedding. In their work, a modified ranking objective function (as illustrated in Figure 2.18) was derived from the WSABIE model [WBU10].

2.3 Related Work in Machine Learning

In this thesis, we design three main frameworks to solve the challenges in attribute learning on image/video understanding. Such frameworks are tightly related to many well studied machine learning algorithms. Specifically,

- the probabilistic topic model inspired the framework in Chapter 3;
- the framework in Chapter 4 is based on graph-based label propagation and Canonical Component Analysis, and the projection domain shift problem is also related to the problems of domain adaptation in the machine learning community;
- the related work of robust ranking and robust learning to rank from crowdsourced pairwise annotations is reviewed and compared with the algorithms in Chapter 5.

So in this section, we briefly review these works in machine learning.

2.3.1 Probabilistic Topic Model

Probabilistic topic models (PTMs) (Blei *et al.* [BNJ03]) have been used extensively in modeling images (Wang *et al.*[WBL09]) and videos (Wang *et al.* [WM09], and Niebles *et al.* [NWFF08]) via learning a low-dimensional topic representation. PTMs are related to attribute learning in that multiple tags can be modeled generatively [BJ03, WBL09], and classes can be defined in terms of their typical topics [WBL09, BM07, HLGX11, HGX11a]. However these topic-representations are generally discovered automatically and lack the semantic meaning which attribute models obtain by supervising the intermediate representation.

There has been limited work (only Yu *et al.* [YA10] and Fu *et al.* [FHXG12]) using topics to directly represent attributes, and provide attractive properties of attribute learning such as zero-shot learning. These are limited to user-defined attributes only [YA10], or formulated in a computationally non-scalable way and for a single modality only [YA10, FHXG12]. In contrast to [YA10] (as well as most annotation studies [TYH+09, THW+09, SHH+07, QHR+07]), in Chapter 3 we will leverage the ability of topic models to learn unsupervised representations from data; and in contrast to [WM09, WBL09, NWFF08, BM07], their framework also leverages prior knowledge of user-defined classes and attributes. Together, these properties provide a complete and powerful semi-latent semantic attribute-space. Scalability can also be a serious issue for topic models applied to video, as most formulations take time proportional to the volume of features (Wang *et al.* [WBL09]). The unstructured social activity attribute (USAA) dataset is larger than text datasets which have been addressed with large-scale distributed algorithms and supercomputers (Newman *et al.*[NAS09]). Therefore chapter 3 will generalize ideas in sparse equivalence class updating to make inference tractable in M2LATM.

2.3.2 Graph-based Label Propagation

Graph-based label propagation is well-studied for semi-supervised learning problems. In general, graph-based semi-supervised methods define a graph with labelled/unlabelled examples as nodes and the similarity of examples as edges, and assume that the label smoothness over the whole graph, i.e., neighbouring labels tend to have the same label (as shown in Figure 2.20). Thus intrinsically such methods are nonparametric, discriminative and transductive. When distinctive graph types are used, there are three different categories for graph-based label propagation: classification on traditional 2-graphs, multi-graph and Hypergraphs.

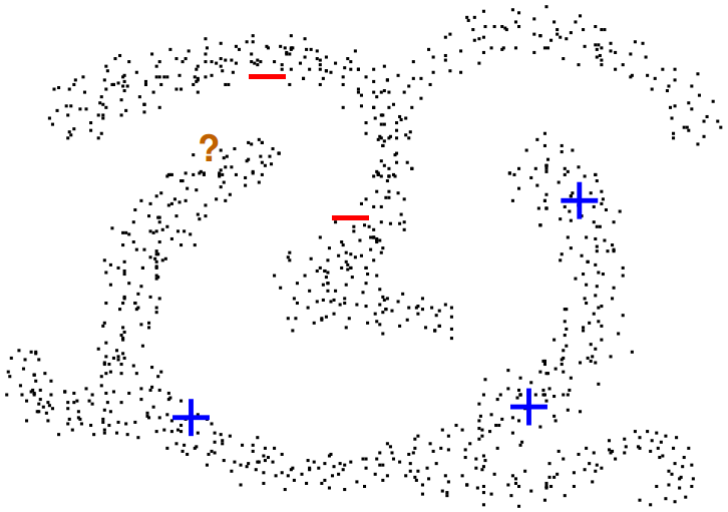


Figure 2.20: Motivations for classification by graph-based label propagation. Images from [Zhu07].

Classification on traditional 2-graphs is the most widely studied for semi-supervised problems (Zhou *et al.* [ZBL+04] and Zhu [Zhu07]). Graph-based label propagation has been used for zero-shot learning in Rohrbach *et al.* [RES13]. They constructed a single graph in the original low-level feature space with the semantic space prototypes used to help label propagation in a heuristic way and no more than two semantic spaces can be used simultaneously.

This thesis focuses on classification on multi-view graphs (C-MG) and Hypergraphs. Most C-MG solutions are based on the seminal work of Zhou *et al.* [ZB07] which generalised spectral clustering from a single graph to multiple graphs by defining a mixture of random walks on multiple graphs. However, crucially, the influence/trustworthiness of each graph is given by a weight that has to be pre-defined and its value has a great effect on the performance of C-MG [ZB07]. In Chapter 4 and [FHX+14a] we extended the C-MG algorithm in [ZB07] by introducing a Bayesian prior weight for each graph, which can be measured automatically from data. Bayesian model averaging was thus applied to fuse multi-view graphs into a single one. Their experiments show that our TMV-HLP algorithm is superior to [ZB07] and [RES13].

Classification on Hypergraphs is a generalisation of classification on traditional 2-graph and C-MG. Hypergraphs have been used as an effective tool to align multiple data/feature modalities in data mining [LHS+13], multimedia [FGZ+10] and computer vision [LLS+13, HYLC13] applications. Since a hypergraph is the generalisation of a traditional 2-graph with each hyperedge connecting a set of nodes (vertices), it can better cope with noisy nodes and thus achieve

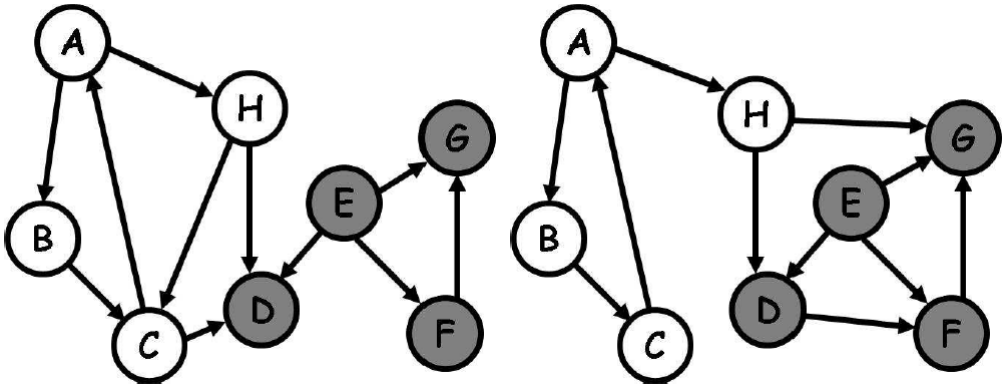


Figure 2.21: Classification with multiple graphs in [ZB07]. There are two directed graphs over the same set of vertices. Those vertices belong to two different classes respectively denoted by gray and white circles.

better performance than pairwise 2-graphs [HLM09, HLZM10, FGZ+10]. Fu *et al.* [FGZ+10] proposed to construct a spatio-temporal shot hypergraph for multi-view video summarisation. Such a hypergraph was used to systematically model the complex multi-view video correlations in which each graph node denotes a video shot, while each type of hyperedge characterizes the relationship among shots. No existing work considers employing hypergraphs for multi-view data modelling, except Hong *et al.* [HYLC13].

However, different from the multi-view hypergraphs proposed in Hong *et al.* [HYLC13] which are homogeneous hypergraphs, that is, constructed in each view independently, Chapter 4 constructs a multi-view heterogeneous hypergraph: using the nodes from one view as query nodes to compute hyperedges in another view. This novel graph structure enables better exploitation of the complementarity of different views in the common embedding space which is validated in their experiments.

2.3.3 Canonical Component Analysis (CCA) for Semantic Embedding

A number of successful applications to learning a semantic embedding space reply on Canonical Component Analysis (CCA) (Hotelling *et al.* [Hot36]). Hardoon *et al.* [HSST04] proposed a general method of kernel CCA to learn semantic embedding to web images and their associated text. Such embedding enables a direct comparison between text and images.

Recently the idea of relating low-level feature and semantic views of data has been exploited in visual recognition and cross-modal retrieval. There are many existing works [SFF10, GKIL13,

HG11, WG07] that focused on modelling the images/videos with associated text (e.g. tags on Flickr/YouTube). Multi-view CCA is often exploited to provide unsupervised fusion of different modalities. Gong *et al.* [GKIL13] also investigated the problem of modeling Internet images and associated text or tags and proposed a three-view CCA embedding framework for retrieval tasks. Due to combining more semantic views, the performance of their framework outperformed a number of two-view baselines on the retrieval tasks. Chapter 4 further extends the three-view to much more views of multi-view CCA embedding framework for zero-shot learning problems. Different from most of previous work, their embedding space was transductive. That is, learned from unlabelled target data from which all semantic views are estimated by projection rather than being the original views. This can rectify the projection domain shift problem existed in most zero-shot learning problems.

2.3.4 Domain Adaptation

Domain adaptation methods attempt to address the domain shift problems that occur when the assumption that the source and target instances are drawn from the same distribution is violated. Methods have been derived for both classification (Fernando *et al.* [FHST13]) and regression (Storkey *et al.* [SS07]), and both with (Duan *et al.* [DTXM09]) and without (Fernando *et al.* [FHST13]) requiring label information in the target task. In contrast, the zero-shot learning problem means that most of supervised domain adaptation methods are irrelevant. Critically, Chapter 4 dicussed the projection domain shift problem that exists in zero-shot recognition. The projection domain shift problem differs from the conventional domain shift problems in that (i) it is in-directly observed in terms of the projection shift rather than the feature distribution shift, and (ii) the source domain classes and target domain classes are completely different and could even be unrelated which renders any efforts to align the two domains directly unfruitful. Chapter 4 thus proposes to rectify the projection domain shift problem in a transductive way and relies on correlating different representations of the unlabelled target data in a multi-view embedding space.

In the context of natural language processing, Blitzer *et. al.* [BFK09] studied the zero-shot learning problem of predicting user-satisfaction ratings across different domains (e.g. book reviews to DVD reviews on Amazon). In their work, views correspond to domains, and for them the tasks for both the source and target domains are the same – estimating user ratings, whilst we aim to recognise a different set of object classes in the target domain.

2.3.5 Robust Ranking and Robust Learning to Rank from Crowdsourced Pairwise Annotations

The inference of a global ranking from a population of partial orders, e.g. paired comparisons, has been widely studied in economics [Arr63], statistics [Dav88, JLYY11], and computer science [CK10, XHJ+12, Joa02]. By aggregating pairwise ranking into global ranking, statistical ranking has the potential to be robust against local ranking noise.

Recently, large-scale pairwise annotations in computer vision are increasingly collected as human intelligence tasks (HIT) using crowdsourcing services, e.g. AMT (Amazon Mechanical Turk). Many studies [KCS08, SF08, PH12] highlighted the necessity of validating random or malicious labels/workers and gave some filtering heuristics for outliers during data collection. However, existing approaches to annotation noise are primarily based on majority voting where each instance is annotated multiple times and averaged in order to remove outliers. Majority voting was shown to achieve 'expert' level of labelling quality within a natural language processing context [SOJN08, CB09]. However, it requires a costly volume of redundant annotations. Moreover, majority voting for pairwise comparison data is only a local (per-pair) inconsistency filtering method; it thus has no effect on global inconsistency and even risks introducing additional inconsistency [Geh83].

HodgeRank has been proposed in [JLYY11] for relatively sparse graphs to decompose each edge flow into three orthogonal components, the gradient flow representing the L_2-optimal global ranking and two cyclic flows for measuring the local and global inconsistencies respectively. Later, Hodgerank has been successfully applied to subjective video quality assessment [XHJ+12, XHY12]. In particular, HodgeRank provides some diagnostic information about outliers. For the purpose of robust ranking the crowdsourced pairs, Xu *et al.* [XXHY13] proposed to assess Quality of Experience (QoE) using crowdsourced pairwise comparison a robust rating scheme derived from Huber-LASSO which is based on robust regression with Huber's Loss [Hub81] and Hodge Decomposition on graphs (Jiang *et al.* [JLYY11]). Yu *et al.* [Yu12] proposed an angular embedding model that maps pairwise comparisons onto a circle and finds the global ranking score via a primary eigenvector solution in the presence of noise. It had been employed to solve a number of vision problems including image denoising (Yu [Yu12]) and object segmentation (Maire *et al.* [MYP11]). Outliers can be also explicitly identified using Transitivity Satisfaction Rate (TSR) (Chen *et al.* [CWCL09b]).

To learn ranking functions for applications such as interestingness prediction [FHX+14b] and relative attribute prediction [PG11b, KPG12], a feature representation of the data points must be used as model input in addition to the local ranking orders. This is addressed in learning to rank which is widely studied in machine learning community [CQL+07, LGL+08, SQTW09, CECC08]. Critically, in practice outliers are still a big issue for learning from crowdsourced pairwise annotations. In Chapter 5, we will extend the Huber-LASSO framework in robust ranking and further proposed a Unified Robust Learning to Rank (URLR) framework for robust learning to rank from crowdsourced annotations. We showed theoretically and experimentally that by jointly solving the outlier detection and ranking estimation problems, the framework achieved better outlier detection than existing statistical ranking methods and better ranking prediction than existing learning to rank method such as rankSVM without outlier detection.

2.4 Summary

The preceding discussions have covered essential issues and studies in the literature regarding attribute learning for image and video understanding. Some widely used attribute learning models are introduced. We particularly discuss and compare binary vs. relative attributes, user-defined vs. data-driven attributes, image vs. video attributes, as well as the low-level features and datasets. We also review other semantic representations beyond attributes, and machine learning work related to this thesis.

The existing methods have shown promising results of attribute learning for image and video understanding. Nevertheless, there are still several open problems and limitations that they do not solve. Firstly, the user-defined attributes are very limited in analysing complex image and video data. The user-defined attributes are defined by extra-knowledge of either expert users or a concept ontology. Thus these attributes are affected intrinsically by *sparse*, *incomplete* and *ambiguous* annotations. Secondly, the existing attribute learning models suffer from the projection domain-shift problems, prototype sparsity problems and inability to combine multiple semantic representations. Thirdly, how to learn from noisy annotations of relative attributes is still an unsolved problem.

In the subsequent chapters of this thesis, our approach is formulated to address these limitations by the following approach: learning latent attributes in Chapter 3 to break the limitations of user-defined attributes; transductive multi-view embedding in Chapter 4 to tackle the problems

of projection domain-shift, prototype sparsity and the inability to combine multiple semantic representations; robust learning of relative attributes in Chapter 4 to learn from noisy annotations of relative attributes.

Chapter 3

Learning Latent Attributes

In this Chapter, we are interested in automatic classification and annotation of unstructured group social activities and complex image classes. Particularly, we focus on home videos of social occassions such as graduation ceremony, birthday party, and wedding reception in USSA dataset of Chapter 2.1.6.6 which feature activities of group of people ranging anything between a handful to hundreds (Fig. 1). By classification, we aim to categorise each video/image into a class; and by annotation we aim to predict what are present in the video/image. This implies a wide range of multi-modal annotation types including object (e.g. group of people, cake, balloon), action (e.g. clapping hands, hugging, taking photos), scene (e.g. indoor, garden, street), and sound (e.g. birthday song, dancing music). We consider that the problems of classification and annotation are inter-related and should be tackled together.

We propose to solve the problems using an attribute learning framework, where annotation becomes the problem of attribute prediction and image/video classification is helped by a learned attribute model. Attributes describe the characterisitics that embody an instance or a class. Essentially attributes answer the question of describing a class or instance in contrast to the typical (classification) question of naming an instance. The attribute description of an instance or category is useful as a semantically meaningful intermediate representation to bridge the gap between low level features and high level classes. Attributes thus facilitate transfer and zero-shot learning to alleviate issues of the lack of labelled training data, by expressing classes in terms of well known attributes.

Learning user-defined attributes is an effective way for transfer learning tasks. Nevertheless,

the user-defined attributes may be limited when used to explore complex multi-modal visual data, since these attributes are defined by extra knowledge from either user experts or concept ontologies and the definition process has no direct linkage with the visual recognition tasks. The possibly poor annotation quality of user-defined attributes may further negatively affect attribute learning algorithms. In most cases, the annotations of user-defined attributes are *sparse*, *incomplete* and *ambiguous*.

These problems are particularly prominent when we apply attribute learning to understand complex consumer videos. The visual data of consumer videos are of unstructured social group activity, i.e. an unconstrained space of objects, events and interactions. The casual nature of this data makes it difficult to extract good features, since they are typically captured with low resolution, poor lighting, occlusion, clutter, camera shake and background noise.

To this end, we propose a framework which can jointly learn user-defined and latent attributes. This chapter systematically formulates a semi-latent attribute space learning framework of learning multi-modal user-defined and latent attributes for automatic classification and annotation of unstructured group social activity. In contrast to existing work of attribute learning for image object class or simple human action classification, this work for the first time, tackles the problem of attribute learning for understanding group social activities with sparse and incomplete labels. In particular we focus on videos of social group activities, which are particularly challenging and topical examples of this task because of their multi-modal content and complex and unstructured nature relative to the density of annotations.

The main content of this Chapter has been previously published in

1. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong; "Attribute Learning for Under- standing Unstructured Social Activity", European Conference on Computer Vision (ECCV) 2012;

2. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong "Learning Multi-modal Latent At- tributes" IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI), 36(2), 303-316, Feb 2014;

3.1 Problem Context and Definition

We first formally introduce the problem of attribute-based learning before developing our contributions in the next section. Learning to detect or classify can be formalised as learning a mapping $F : \mathcal{X}^d \rightarrow \mathcal{Z}$ of d-dimensional raw data $\mathcal{X}$ to label $\mathcal{Z}$ from training data $D = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$. A variant of the standard approach considers a composition of two mappings [PHPM09]:

$$F \quad = \quad S(L(\cdot)), L : \mathcal{X}^d \rightarrow \mathcal{Y}^p, S : \mathcal{Y}^p \rightarrow \mathcal{Z}, \tag{3.1}$$

where L maps the raw data to an intermediate representation $\mathcal{Y}^p$ (typically with $p \ll d$) and then S maps the intermediate representation to the final class $\mathcal{Z}$. Examples of this approach include dimensionality-reduction via PCA (where L is chosen to explain the variance of $\mathbf{x}$ and $\mathcal{Y}^p$ is the space of orthogonal principal components of $\mathbf{x}$) or linear discriminant and multi-layer neural networks (where L is optimised to predict $\mathcal{Z}$).

Attribute learning [LNH09, PHPM09] exploits the idea of requiring $\mathcal{Y}^p$ to be a *semantic attribute* space. L and S are then learned by direct supervision with instance, attribute vector and class tuples $D = \{(\mathbf{x}_i, \mathbf{y}_i, z_i)_{i=1}^n\}$. This has benefits for sparse data learning including multi-task, N-shot and zero-shot. In multi-task learning [STT11] the statistical strength of the whole dataset can be shared to learn L, even if only subsets corresponding to particular classes can be used to learn each class in S. In N-shot transfer learning, the mapping L is first learned on a large "source/auxiliary" dataset D. We can then effectively learn a much smaller "target" dataset $D^* = \{(\mathbf{x}_i, z_i^*)\}_{i=1}^m$, $m \ll n$ containing novel classes z^* by transferring the attribute mapping L to the target task, leaving only parameters of S to be learned from the new dataset D^*. The key unique feature of attribute learning is that it allows zero-shot learning: the recognition of novel classes without any training examples $F : \mathcal{X}^d \rightarrow \mathcal{Z}^*$ ($\mathcal{Z}^* \notin \mathcal{Z}$) via the learned attribute mapping L and a manually specified attribute description S^* of the novel class.

3.2 Semi-latent Semantic Attribute Space

Most prior attribute learning work [FEH10, FEHF09, LNH09, KBBN09] unrealistically assumes that the attribute space $\mathcal{Y}^p$ is completely defined in advance, and contains sufficiently many attributes which are both *reliably detectable* from $\mathcal{X}$ and *discriminative* for $\mathcal{Z}$. We now relax these assumptions by performing *semantic feature reduction* [PHPM09] from the raw data to a lower dimensional *semi-latent semantic attribute space* (illustrated in Fig. 1.5(b)).

Semi-latent semantic attribute space: *A p dimensional metric space where p_{ud} dimensions encode manually specified semantic properties, and p_{la} dimensions encode latent semantic properties determined by some objective given the manually defined dimensions.*

We aim to define an attribute-learning model L which can learn a semi-latent attribute space from training data D where $|\mathbf{y}| = p_{ud}$, $0 \leq p_{ud} \leq p$. That is, only a p_{ud} sized subset of the attribute dimensions are user-defined, and p_{la} other relevant latent dimensions are discovered automatically. The attribute-space is thus partitioned into observed and latent subspaces: $\mathcal{Y}^p = \mathcal{Y}_{ud}^{p_{ud}} \times \mathcal{Y}_{la}^{p_{la}}$ with $p = p_{ud} + p_{la}$. To support a full spectrum of applications, the model should allow:

1. an exhaustively and correctly specified attribute space $p = p_{ud}$ (corresponding to previous attribute learning work);

2. a partially known attribute space $p = p_{ud} + p_{la}$ (corresponding to an incomplete ontology);

3. a completely unknown attribute space $p = p_{la}$. Such a model would go beyond existing approaches to bridge the gap (Fig. 1.5(a)) between exhaustive and unspecified attribute ontologies. As we will see, performing classification in this semi-latent space will provide increased robustness to the amount of domain-knowledge/ontology creation budget, and to annotation noise as compared to conventional approaches.

3.3 Multi-modal Latent Attribute Topic Model

To learn a suitable attribute model L (Eq. (3.1)) with the flexible properties outlined in the previous section, we will build on probabilistic topic models [BNJ03, HLGX11]. Essentially we will represent each attribute with one or more topics, and add different types of constraints to the topics such that some topics will represent user-defined attributes, and others latent attributes.

First, we briefly review the standard Latent Dirichlet Allocation (LDA) [BNJ03] approach to topic modeling. Applied to video understanding [HGX11b, HLGX11, HGX11a, NWFF08], conventional LDA learns a generative model of videos $\mathbf{x}_i$. Each quantized feature x_{ij} in clip i is distributed according to a discrete distribution $p(x_{ij}|\beta_{y_{ij}}, y_{ij})$ with a Dirichlet parameter β corresponding to its (unknown) parent topic y_{ij}. Topics in video i are distributed according to

another discrete distribution $p(\mathbf{y}_i|\theta_i)$ paramaterized by the Dirichlet variable θ_i. Finally, the prior probability of topics in a video are distributed according to the $p(\theta_i|\{\alpha\})$ with parameter α.

Standard LDA is uni-modal and unsupervised. Unsupervised LDA topics can potentially *represent* fully latent (GF) attributes. We will modify LDA to constrain a subset of topics (UD and CC) to *represent* conventional supervised attributes [LNH09, PHPM09]. The three attribute types are thus given a concrete representation in practice by a single topic model with three types of topics (UD, CC and GF), differing in terms of the constraints with which they are learned. We next detail our M2LATM including learning from (1) supervised attribute annotations and (2) multiple modalities of observation.

3.3.1 Attribute-topic Model

In order to model supervised user-defined attribute annotations, M2LATM establishes a topic-attribute correspondence so that attribute k is represented by topic k. We encode the (user-defined) attribute annotation for video i via a per-instance vector topic prior α_i. An attribute k is encoded as absent via setting $\alpha_{ik} = 0$, or present via $\alpha_{ik} = 1$. The full joint distribution for a database D of videos with attribute annotations $\{\alpha_i\}$ is:

$$p(D|\{\alpha\},\beta) = \prod_i \int \left(\prod_j \sum_{y_{ij}} p(x_{ij}|y_{ij},\beta) p(y_{ij}|\theta_i) \right) p(\theta_i|\alpha_i) d\theta_i, \tag{3.2}$$

To infer the attributes for a clip, we require the posterior $p(\theta_i, \mathbf{y}_i|\mathbf{x}_i, \alpha_i, \beta)$. As for LDA [BNJ03], this is intractable to compute exactly. Variational inference approximates the full posterior with a factored variational distribution:

$$q(\theta_i, \mathbf{y}_i|\gamma_i, \phi_i) = q(\theta_i|\gamma_i) \prod_j q(y_{ij}|\phi_{ij}). \tag{3.3}$$

where γ_{ik} parameterizes the Dirichlet factor of topic/attribute k proportions θ_i within clip i; and ϕ_{ijk} parameterizes the discrete posterior y_{ij} of topic/attributes for feature x_{ij}. Optimizing the variational bound results in the updates:

$$\begin{aligned} \phi_{ijk} &\propto \beta_{x_{ijk}} \exp(\Psi(\gamma_{ik})), \\ \gamma_{ik} &= \alpha_{ik} + \sum_j \phi_{ijk}, \end{aligned} \tag{3.4}$$

where Ψ is the digamma function. Iterating Eq. (3.4) to convergence completes the variational E-step of an expectation maximisation (EM) algorithm. The M-step updates parameter β by

maximum likelihood: $\beta_{vk} \propto \sum_{i,j} \mathbf{I}(x_{ij} = v)\phi_{ijk}$. After EM learning, each attribute/topic y (e.g., clapping hands or singing) will be associated with a particular subset of the low-level features via $p(x|y,\beta)$ and learned parameter β.

3.3.2 Learning Multiple Modalities

Topic model generalizations exist to jointly model multiple translations of the same text [MWN+09] via a common topic profile θ, where one language could be considered one modality. However, this is insufficient because as we have discussed, a given attribute may be unique to a particular modality. To model multi-modal data $D = \{D_m\}_{m=1}^{M}, D_m = \{\mathbf{x}_{im}\}$, we therefore exploit a unique topic prior θ_m per-modality m as follows:

$$\begin{aligned} p(\{D_m\}|\{\alpha\},\{\beta_m\}) &= \prod_{i,m} \int d\theta_{im} p(\theta_{im}|\alpha_i) \\ &\times \left(\prod_j \sum_{y_{ijm}} p(x_{ijm}|y_{ijm},\beta_m) p(y_{ijm}|\theta_{im}) \right). \end{aligned} \tag{3.5}$$

By sharing the annotations α across modalities, but allowing a unique per-modality prior θ_m, the model is able to represent both attributes with strong multi-modal correlates (e.g., clapping hands) and those more unique to a particular modality (e.g., laughter, candles). Moreover, this approach provides an automatic way to deal with different modalities being expressed on different scales. Different scale modalities is a serious problem for most topic models hoping to simply concatenate multi-modal data: either one modality dominates or words underflow is risked if data is normalized. For example, 99% of the feature frequencies in USAA are in the range of $[0,8]$ (appearance), $[0,450]$ (motion), and $[0,50]$ (auditory). For this reason studies [YA10] often only use a single modality when many are available. Fig. 3.1 provides a graphical model representation of M2LATM.

3.3.3 Learning User-defined and Latent Attributes

With no user-defined attributes ($p = p_{la}, p_{ud} = 0$), a p-topic LDA model provides a mapping L from raw data $\mathbf{x}$ to a p-dimensional latent space by way of the variational posterior $q(\theta|\gamma)$. This is a discrete analogy to the common use of PCA to reduce the dimension of continuous data. However, to (i) support user-defined attributes when available and (ii) ensure the latent representation is discriminative, we add constraints.

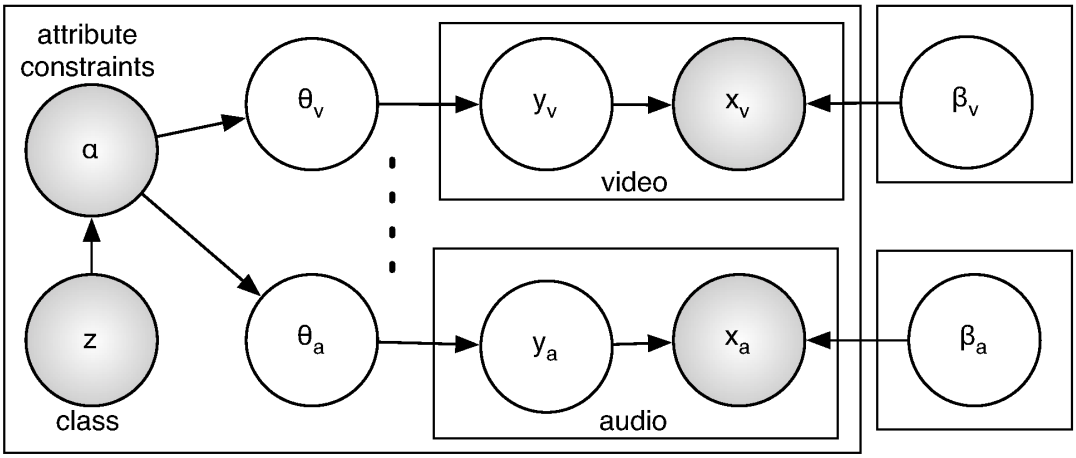


Figure 3.1: Graphical model for M2LATM.

User-defined attributes are typically provided in terms of length p^{ud} binary vectors $\mathbf{v}^{ud}$ specifying the attributes of class z or instance i [LNH09, PHPM09]. We have no prior knowledge of the relation between $\mathbf{v}^{ud}$ and each word (i, j), so $\mathbf{v}^{ud}$ cannot determine $\mathbf{y}$ directly. To enforce the user-defined attribute constraint, we define a *per instance* prior $\alpha_i = [\alpha_i^{ud}, \alpha_i^{la}]$, setting $\alpha_{i,k}^{ud} = 0$ if $v_{i,k}^{ud} = 0$ and $\alpha_{i,k}^{ud} = 1$ otherwise. It enforces that instances i lacking an attribute k can never use that attribute to explain the data; but otherwise leaving the model to infer attribute proportions, modality and word correspondence.

To learn the latent portion of the attribute-space, we could simply leave the remaining portion α^{la} of the prior unconstrained. However, for the latent space to be useful, it should be both *discriminative* (for class) and *generalizable* (to potential new classes) [HLGX11, HGX11a]. To obtain both of these properties, we split the prior into components for "class-conditional" (CC) and "generalized free" (GF) topics. When learned jointly with UD attributes and with appropriate constraints, CC topics will be selective for known classes and GF topics will represent attributes shared between known classes, and hence likely to generalize. Specifically, we split the latent space prior $\alpha_i^{la} = [\alpha_i^{cc}, \alpha^{gf}]$. In the CC component $\alpha_i^{cc} = \{\alpha_{i,z}\}_{z=1}^{N_z}$, each subset $\alpha_{i,z}$ corresponds to a class z. For an instance i with label z_i, set $\alpha_{i,z=z_i}^{cc} = 1$ and all other $\alpha_{i,z\neq z_i}^{cc} = 0$. This enforces that only instances with class z can allocate topics y_z^{cc} and hence that these topics are discriminative for class z. The GF component of the latent space prior is uniform $\alpha^{gf} = 1$, meaning that GF topics are shared between all classes and thus represent aspects shared among all the data.

3.3.4 Classification

To use M2LATM for classification, we define the mapping L in Eq. (3.2) as the posterior statistic γ in Eq. (3.9). The remaining component to define is the attribute-class mapping S. Importantly, for our complex data, this mapping is not deterministic (i.e., 1:1 correspondence between attributes and classes assumed in [LNH09, PHPM09]). Like [LKS11], we therefore use standard classifiers to learn this mapping from (z_i, γ_i) pairs obtained from our M2LATM attribute learner.

3.3.5 Surprising Attributes

M2LATM can also be used to find videos which exhibit surprising/abnormal semantics. Given the training labels and estimated set of posterior semi-latent topic profiles $\{z_i, \gamma_i\}$, we can fit a multi-variate Gaussian $\mathcal{N}(\mu_\gamma^z, \Sigma_\gamma^z)$ to the profile of examples from each class z. At test time, once the class z^* of a given instance is estimated, we can detect surprising attribute semantics by computing the likelihood $p(\gamma^*|\mu_\gamma^{z^*}, \Sigma_\gamma^{z^*})$. Importantly, unlike earlier notions of attribute-surprise [FEHF09], this approach (i) also considers surprising latent attributes, and (ii) inter-attribute and inter-modality correlations.

3.4 Semi-latent Zero Shot Learning and Inference

3.4.1 Semi-latent Zero Shot Learning

Zero-shot learning addresses classification of unseen classes via semantic attribute descriptions rather than via learning from training examples. A description $\mathbf{v}_{z^*}^{ud} \in \mathcal{Y}_{ud}$ for a new class z^* is provided in terms of attributes from human prior knowledge. Existing approaches [LNH09, PHPM09] define simple deterministic prototypes $\mathbf{v}_{z^*}^{ud}$ in terms of UD attributes only, and classify by matching these templates $\mathbf{v}_{z^*}^{ud}$ to the estimated UD attributes for each test instance, e.g., by nearest-neighbour (NN) [FEHF09] or naïve-Bayes. Using NN, conventional zero-shot classification of test instance $\mathbf{x}^*$ with UD attribute representation $\mathbf{y}^{*,ud}$ is:

$$f(\mathbf{x}^*) \quad = \quad \arg\min_{z^*} \{ \left\| \mathbf{y}^{*,ud} - \mathbf{v}_{z^*}^{ud} \right\| \}. \tag{3.6}$$

However, in this approach one needs a large ontology of attributes, and to specify an (impractically long) definition of each new class in terms of every attribute in the ontology. Counter-intuitively, we can work with a smaller UD ontology ($p_{ud} \geq 1$) and leverage the latent portion of the attribute-space to still obtain a rich representation for classification. We project a

short/incomplete UD attribute description of a novel class into the complete semi-latent attribute space description as follows:

1. Input a test set $D^* = \{\mathbf{x}^*\}$ containing novel classes, and UD attribute prototypes $\mathbf{v}_{z^*}^{ud}$ for those classes.

2. Infer attributes $\mathbf{y}^* = [\mathbf{y}^{*,ud}, \mathbf{y}^{*,la}]$ for each test data $\mathbf{x}^*$ (given by γ in Eq. (3.4))

3. Let $NN_k^{ud}(\mathbf{v}_{z^*}^{ud}, \{\mathbf{y}^{*,ud}\})$ denote the set of k nearest UD neighbours in D^* to each prototype $\mathbf{v}_{z^*}^{ud}$ in $\mathcal{Y}_{ud}$.

4. Project UD prototypes $\mathbf{v}_{z^*}^{ud} \in \mathcal{Y}_{ud}$ into the full attribute space $\mathcal{Y}$ by averaging their nearest neighbours (Eq. (3.7)).

5. Perform zero-shot classification in the full attribute space $\mathcal{Y}$ (Eq. (3.8)).

$$\mathbf{v}_{z^*} = \frac{1}{k} \sum_{\mathbf{y} \in NN_k^{ud}(\mathbf{v}_{z^*}^{ud}, D^*)} \mathbf{y}, \tag{3.7}$$

$$f(\mathbf{x}^*) = \arg\min_{z^*} \|\mathbf{y}^* - \mathbf{v}_{z^*}\|. \tag{3.8}$$

The mechanism of this algorithm is schematically illustrated in two dimensions by Fig. 3.2. The one dimensional UD prototype $y^{*,ud}$ (blue line) only weakly identifies (shading) the target class 'x'. After projecting into the full space, the two-dimensional prototype (blue dot) more clearly identifies (shading) the target class.

Our approach can be viewed in a few ways: as transductively exploiting the test data distribution; or as one iteration of an EM-style algorithm for data with partially-known parameters and unknown variables (in contrast to the typical semi-supervised learning case of partially known variables and unknown parameters [Zhu07]). Previous ZSL studies are constrained to user-defined attributes, thus being critically dependent on the completeness of the user attribute-space. In contrast, our approach uniquely leverages a potentially much larger body of latent attributes via a loose manual definition of a novel class. We will show later this approach can significantly improve zero-shot learning performance.

3.4.2 Efficient Variational Inference and Implementation

Our formulation thus far, as well as the earlier work [FHXG12] and LDA in general, infers the posterior over topics/attributes *for each word* (i.e. Eq. (3.4) indexed by word j). This is true

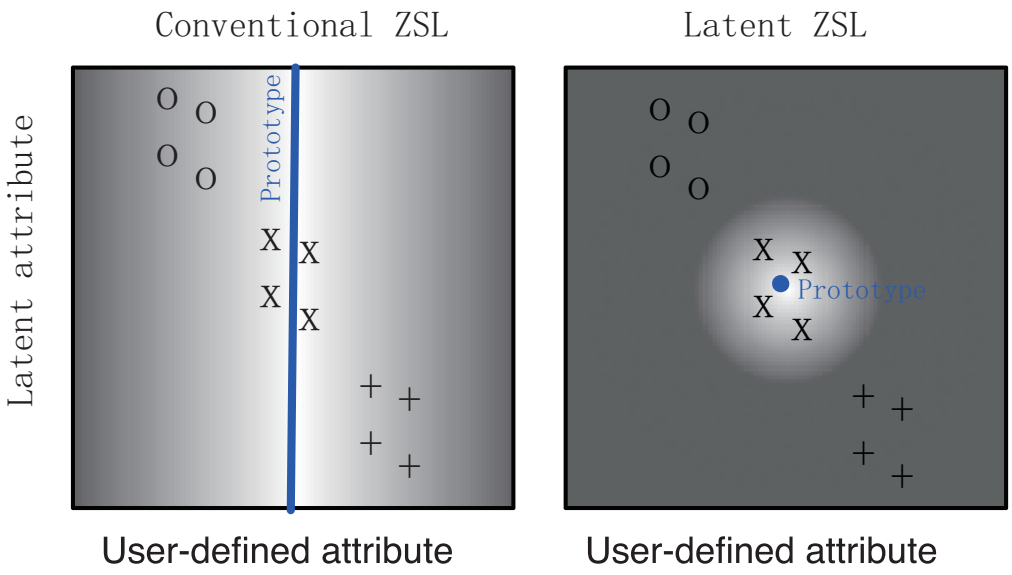


Figure 3.2: Schematic illustration of latent ZSL mechanism.

whether solved with variational inference [BNJ03] or MCMC [YA10, MWN+09]. Computation is thus $\mathcal{O}(NK)$ for N total words and K topics. For our video dataset, where words correspond to dense interest point detections, N is of the order 10^{10} and grows with video length. Conventional topic models do not scale to this data in either processing or memory demands, requiring days to run on in practice. In contrast, approaches such as support vector machines (SVM) [JYC+11] use the same data, but operate on word proportions within the vocabulary V. SVMs are thus $\mathcal{O}(V)$ and therefore significantly faster than conventional $\mathcal{O}(N)$ topic models because typically V is $\leq 10^4$.

Inspired by [AWST09], we observe that while each observation x_{ijm} has an associated topic posterior, all instances of the same vocabulary item $x \in V$ within one video have the same posterior ϕ. Exploiting this equivalence class, the same inference can therefore be expressed in the $\mathcal{O}(V)$ vocabulary domain, rather than the $\mathcal{O}(N)$ word domain. Inference for multiple modalities m expressed in vocabulary-domain is thus:

$$\begin{aligned} \phi_{ivkm} &\propto \beta_{vkm} \exp(\Psi(\gamma_{ikm})), \\ \gamma_{ikm} &= \alpha_{ik} + \sum_v \mathbf{h}_v(\mathbf{x}_{im}) \phi_{ivkm}. \end{aligned} \tag{3.9}$$

Here, $\mathbf{h}_v(\mathbf{x}_{im})$ denotes the histogram of observations in $\mathbf{x}_{im}$, and the topic posterior matrix $\phi_{x\cdot\cdot m}$ is now of size VK instead of NK. Further efficiencies may be obtained by observing that only sufficient statistics for vocabulary elements observed in each document need to be computed. That is, Eq. (3.9) can be updated as a sparse matrix operation for unique observations U_i at

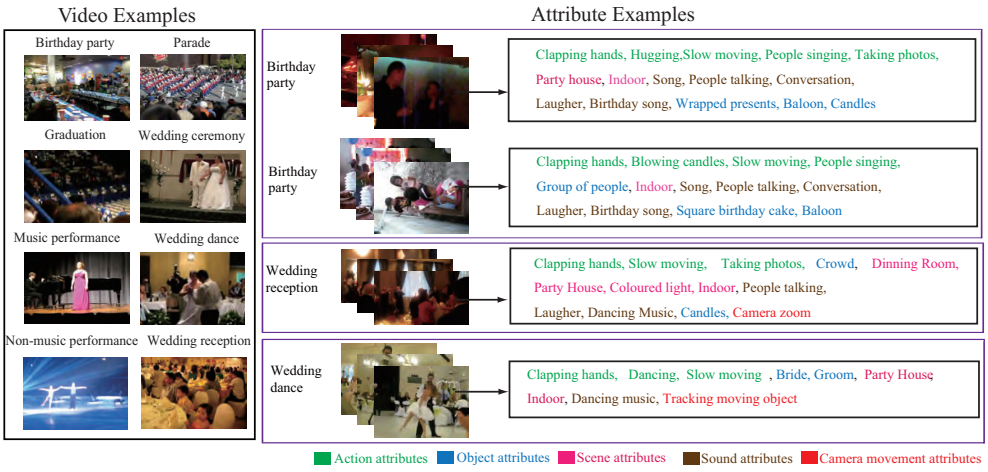


Figure 3.3: Examples from the eighth classes and video attributes in the USAA dataset. Different types of attributes in visual and auditory modalities are shown in different color.

$\mathcal{O}(U_i K)$ cost per document i, where typically $U_i \ll V \ll N$.

3.5 Experiments

We first introduce our datasets and baseline models (Sections 3.5.1 and 3.5.3), then report quantitative results obtained for the three main sparse data learning problems: multi-task learning, N-shot learning and zero-shot learning (Sections 3.5.4 and 3.5.5). We also perform additional analysis on attribute-understanding tasks, robustness, and computation time (Sections 3.5.6-3.5.9).

3.5.1 Unstructured Social Activity Attribute Dataset

In previous work [FHXG12], we introduced a new attribute dataset for social activity video classification and annotation: *unstructured social activity attribute* (USAA)[1]. We selected 100 videos per-class for training and testing from 8 classes of social activities in the CCV dataset [JYC+11] (thus 1600 videos in total). We defined a wide variety of relevant attributes (illustrated in Fig. 3.3), and manually annotated their ground truth at the individual video level. The classes were selected as the most complex social group activities and the video length ranged from 20 seconds to 8 minutes. The eight classes are: birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception.

We experimented with two attribute-ontologies. In the first ontology, we extracted keywords from the CCV class definitions [JYC+11] and used these to obtain a set of 15 attributes. For

[1] http://www.eecs.qmul.ac.uk/~yf300/USAA/download/

example, the definition of graduation party is: "Graduation ceremony with *crowd*, or one or more people wearing *graduation caps* and gowns", from which we obtain attributes "crowd" and "graduation cap". In order to obtain a more exhaustive ontology of attributes, we further annotated a total of 69 attributes covering every conceivable property for this dataset including actions, objects, scenes, sounds, and camera movement. These attributes are illustrated in Figure 3.3. Real-world video will rarely contain such extensive tagging. However, this exhaustive annotation gives the freedom to learn on various subsets in order to quantify the effect of annotation density and biases.

Using the 69 ground-truth attributes (average density 11 per video) directly as input to a SVM, the videos can be classified with 86.9% accuracy. Individual SVM-attribute detectors achieve the mean average precision in the range $[0.22, 1]$ with average 0.785 across the entire ontology. The high variability reflects some attributes which can be detected almost perfectly (e.g., indoor scene), and others which cannot be detected given the available features (e.g., parade float). These points illustrate the challenge of these data: there is sufficient intra-class variability that even perfect knowledge of the attributes instance is insufficient for perfect classification; and moreover many attributes cannot be detected reliably.

3.5.2 Video Feature Extraction and Representation

The foundation for video content understanding is extracting and representing suitably informative and robust features. This is especially challenging for unconstrained consumer video and unstructured social activity due to dramatic within-class variations, as well as noise sources of occlusion, clutter, poor lighting, camera shake and background noise [JYN10]. Global features provide limited invariance to these noise-sources. Local keypoint features collected into a bag-of-words (BoW) are considered state of the art [JYC$^+$11, JYN10, YHWZ11]. We follow [JYC$^+$11, JYN10, YHWZ11], in extracting features for three modalities, namely static appearance, motion, and auditory. Specifically, we employ scale-invariant feature transform (SIFT) [Low04], spatial-temporal interest points (STIP) [Lap05], and mel-frequency cepstrum (MFCC) respectively. The details of extracting these visual features are discussed in Chapter 2.1.5.

3.5.3 Experiment Settings

For each experiment, we use 100 videos per class for testing, and a set of 100 or fewer per class for training both the attribute detectors and category classifiers. We report test set performance

averaged over 5 cross-validation folds with different random selections of instances, classes, or attributes held out as appropriate. We compare the following models:

- *Direct*: Direct KNN or SVM classification on raw data without attributes. SVM is used for experiments when the number of training instances is bigger than 10; and KNN otherwise. Our experiments show that KNN performed consistently better than SVM until the number of training instances is bigger than 10.

- *DAP*: SVM classifiers learn available UD attributes. Then zero-shot learning (ZSL) by Direct Attribute Prediction (DAP), exactly as described by [LNH09]. It is only applicable to ZSL and deterministic attributes.

- *SVM-UD*: SVM classifiers learn available UD attributes. For N-shot learning, a logistic regression (LR) classifier then learns classes given the attribute classifier outputs. LR is chosen over SVM because it was more robust to sparse data. This is analogous to [FEHF09]. For ZSL the SVM posteriors are matched against the manually specified prototype with NN. This is an obvious generalization of DAP [LNH09] to non-deterministic attributes.

- *SCA*: Topic model from [WBL09]. Learns a generative model for both class label and annotations given latent topics, in contrast to the attribute paradigm of expressing classes *in terms of* annotations/attributes. It only applies to multi-task learning.

- *ST*: Synthetic Transfer [YA10]. A ZSL strategy for attribute topic models: Use the source topic model to synthesize training data for novel target classes, which are then learned conventionally. We use this with our topic model. It only applies to ZSL.

- *M2LATM*: Our M2LATM is learned, then a LR classifier learns classes based on the semi-latent topic profile γ. We use 100 topics in total, with 1 UD topic per UD attribute, 1 latent CC per class, and remaining topics are allocated to GF latent attributes.

For all experiments, we cross-validate the regularisation parameters for SVM and LR. For all SVM models, we use the χ^2 kernel. For M2LATM, the user-defined part of the M2LATM topic profile γ is estimating the same quantities as the SVM attribute classifiers, however the latter are slightly more reliable due to being discriminative classifiers, so we use these in conjunction with the latent topic profile for classification. The significance of this is quantified in Section 3.5.7. For semi-latent ZSL, parameter K (Section 3.4) was fixed to 5% of the instances.

3.5.4 Multi-task Learning

3.5.4.1 M2LATM multi-modal latent attributes enhance multi-task learning of sparse data with incomplete ontology.

When all classes are known in advance, shared attributes provide a mechanism for multi-task learning [STT11]. The statistical strength of data supporting each attribute can be aggregated across its occurrences in all classes.

Table 3.1 summarizes our results. We first consider the simplest upper bound scenario where the data is plentiful (100 instances per class, "100I") and the attributes are exhaustively defined (all 69UD, "A/69"). In this case all the models perform similarly except SCA[WBL09], due to the SCA is enforced by the supervised generatic topic model which is relatively weaker than the other compatitors of discriminatively supervised. Next, we consider the sparse data and the incomplete attribute space scenario of interest, with only 10 instances per class to learn from. Here Direct performs poorly due to insufficient data. Limiting the attributes to a randomly selected seven every trial ("R/7"), SVM-UD performs poorly and our M2LATM outperforms all the others by a large margin. Moreover, SVM-UD cannot be applied with a completely held out attribute-ontology ("N/0"), while M2LATM performance is almost unchanged. With no attribute-ontology "N/0", SCA simplifies to supervised LDA [BM07][2]. Our model is thus able to share statistical strength among attributes (unlike Direct); and unlike SVM-UD, it exploits latent attributes to do so robustly to the completeness of the attribute-space definition.

3.5.4.2 M2LATM improves both best and worst case semantic ontologies.

In order to quantify the effectiveness of each attribute in the ontology we ranked the attributes in terms of a simple selection criteria of their "informativeness" used in text categorization [YP97]: Mutual information with the class (informativeness) times reliability (detection rate;). We then contrast performance between a best and worst case user-defined attribute ontology, by using the top and bottom 10% of UD attributes ("T/7" and "B/7") respectively. SVM-UD loses 14% performance from the best to worst case, whereas our M2LATM model is virtually unchanged. In both cases, M2LATM provides a significant improvement over SVM-UD. SCA [WBL09] performs significantly and consistently worse than the other models because it leverages attributes in a weaker way (as annotations rather than constraints), so we do not consider it further.

[2]We used `http://www.cs.princeton.edu/chongw/slda/`

	Direct	SVM-UD	SCA[WBL09]	M2LATM
100I, A/69	**66.0**	65.7	44.0	65.6
10I, A/69	26.8	40.2	32.2	**40.6**
10I, R/7	26.8	26.4	25.6	**38.3**
10I, N/0	26.8	-	17.3	**40.4**
10I, T/7	26.8	32.4	26.0	**38.3**
10I, B/7	26.8	18.2	26.0	**38.9**

Table 3.1: Multi-task classification performance for USAA. 8 classes, chance = 12.5%. Row labels are I: number of training instances per class, A: all attributes, R: random subset of attributes, N: no attributes, T: top attributes, B: bottom attributes.

3.5.5 Transfer Learning

3.5.5.1 M2LATM multi-modal latent attributes enhance N-shot learning of sparse data.

In N-shot transfer learning, one assumes ample examples of a set of source classes, and sparse (N) examples of a *disjoint* set of target classes. To test this scenario, in each trial we randomly split the 8 classes into two disjoint groups of four source and target classes. We use all the data from the source task to train the attribute models (M2LATM and SVM-UD), and then use these to obtain the attribute profiles for the target task. Using the target task attribute profiles we perform N-shot learning, with the results summarized by Table 3.2. Importantly, the SVM-UD attribute learning approach cannot deal with zero attribute situations, so can provide no benefit over Direct here, while our M2LATM improves significantly over Direct ("N/0"). In addition to drawing random subsets of attributes ("R/7" and "R/34"), we also consider the subset of 15 attributes ("O/15") we obtained from the CCV ontology (Section 3.5.1). Our M2LATM performs comparably or significantly better than Direct and SVM-UD in every case. Importantly M2LATM is robust to the both sparse data (performance > 35% for 1-shot learning), and exhaustiveness of the attribute-space definition (no attribute "N/0" performance within 5% of all attribute "A/69" performance). In contrast, Direct suffers strongly under sparse data 1-shot learning, and SVM-UD suffers with sparse attribute-space (7UD "R/7" performance 12% below all attribute performance). The robust performance of M2LATM is enabled by the semi-latent attribute representation.

	1-shot			5-shot			10-shot		
	Direct	SVM-UD	M2LATM	Direct	SVM-UD	M2LATM	Direct	SVM-UD	M2LATM
N/0	29.0	-	**35.3**	33.6	-	**48.0**	35.7	-	**53.0**
R/7	29.0	30.9	**35.9**	33.6	36.9	**48.0**	35.7	38.7	**52.2**
R/34	29.0	35.0	**36.5**	33.6	44.5	**48.9**	35.7	47.5	**52.8**
O/15	29.0	36.1	**37.7**	33.6	46.8	**49.7**	35.7	50.2	**53.3**
A/69	29.0	**39.1**	38.6	33.6	49.7	**52.1**	35.7	52.5	**56.1**

Table 3.2: N-shot classification performance for USAA dataset (4v4 classes, chance = 25%) .

3.5.5.2 M2LATM multi-modal latent attributes enhances zero-shot learning.

Like N-shot learning, the task is to learn transferrable attributes from a source dataset for use on a disjoint target set. Instead of providing training examples, users manually specify the definition of each novel class in the user-defined attribute space. ZSL is often evaluated in simple situations where classes have unique 1:1 definitions in the attribute-space [LNH09]. For unstructured social data [JYC+11], strong intra-class variability violates this assumption, making evaluation more subtle. To define the novel classes, we take the thresholded mean (as in [FEHF09, FHXG12]) of the attribute profiles for each instance of that class from our ground-truth.

Our results are summarized in Table 3.3. The key observation is that using latent attributes to support the user-defined attributes allows M2LATM to improve on SVM-UD [LNH09], which only uses UD attributes in ZSL. This is a surprising and significant result, because it is not obvious that ZSL from human descriptions should be able to be exploit latent data-driven attributes. Additionally, we compare the synthetic data transfer strategy from [YA10], generating $N = 50$ synthetic data instances per class from the zero-shot definition, and training the classifier based on the learned profiles for these. We found that this underperformed DAP in most cases, and M2LATM in every case. This is unsurprising, because synthetic data adds no truly new information: it is generated from the UD word-topic distributions β, learned from the source dataset. M2LATM already uses β, but additionally exploits latent topics.

	SVM-UD	ST[YA10]	M2LATM
R/7	27.1	18.1	**33.8**
O/15	31.3	36.9	**39.4**
R/34	36.7	30.9	**39.2**
A/69	33.2	31.0	**41.9**

Table 3.3: Zero-shot classification performance (%) for USAA (4v4 classes, chance = 25%).

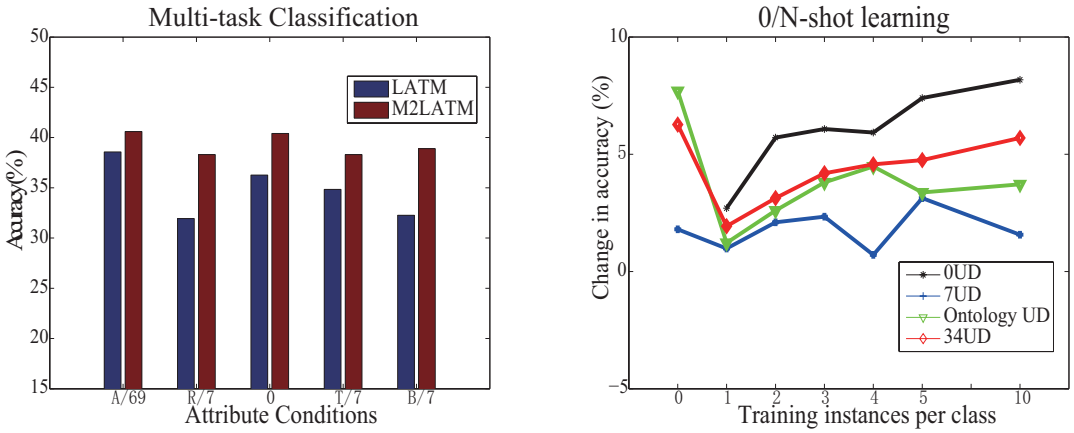


Figure 3.4: Exploiting multi-modality: LATM vs M2LATM for USAA dataset. Left: Multi-task classification. Right: 0/N-shot learning shown as margin of M2LATM over LATM – positive value means increase of accuracy.

3.5.6 Attribute Understanding

3.5.6.1 M2LATM makes effective use of multiple modalities.

An important contribution of M2LATM is explicitly representing the correspondence between attributes and features of each modality, bridging the cross-modal gap. Existing approaches often ignore this issue either by using only one modality [YA10] or taking a weighted average/concatenation [LNH09] of modalities, which introduces issues in selection of scaling/weighting factors. We compare M2LATM against a simpler variant of our approach approach, LATM. LATM takes the standard approach of simply concatenating feature vectors (with rescaling to ensure modalities are represented on the same scale). Explicit multi-modality consistently improves the results relative to simple concatenation in multi-task (Fig. 3.4, left) and transfer (Fig. 3.4, right) learning.

Modality	Attributes
Static (SIFT)	Candles, Dark outdoors, Party House
Motion (STIP)	Slow moving, Crowd, Bright outdoors
Audio (MFCC)	Laughter, Singing, Instrumental music
Static+Motion	Hold microphone, Birthday caps, Crowd
Static+Audio	Singing, People in a row, Fast moving
Audio+Static	Formal speech, Crowds, Dining room

Table 3.4: Top-3 attributes most strongly associated with modalities.

3.5.6.2 M2LATM associates attributes with their observation modality.

To provide further insight into the capabilities of our cross-modal model, we consider a novel task of learning which modalities each attribute appears in. This can be computed from the relative proportion of words assigned by the model to static appearance, motion or auditory modalities when explaining a given topic/attribute. That is, comparing modalities m in $\sum_i \gamma_{ikm}$ for each attribute k. To illustrate this, Table 3.4 reports the top-3 attributes most strongly associated with each modality and each modality pair (as assessed by geometric mean). Clearly most attributes have associations with intuitive modalities.

3.5.6.3 M2LATM can detect semantically surprising multimedia content.

As a final example of attribute understanding, we illustrate some examples of surprising semantics discovered by our framework – based on the correlations encoded in the class-attribute relationships (Section 3.3.5). Fig. 3.5(A) is correctly classified as a birthday party. However, both the *"instrumental music*" (auditory) and "*musical instruments*" (static appearance) attributes are detected (a person sings "happy birthday" using a guitar), which are unusual in birthday party settings. Fig. 3.5(B) is a music performance video, which unexpectedly has the *"costume*" attribute, as there are also costumed actors on stage. A wedding ceremony is shown in Fig. 3.5(C), where guests are unusually drinking during the ceremony ("*drinking glass*" attribute). Fig. 3.5(D) illustrates an example of expected attributes which are surprisingly absent. In this case the video is correctly classified as a parade, however the expected attributes "bright outdoor scene" and "parade float" are absent because it is, unusually, an indoor parade.

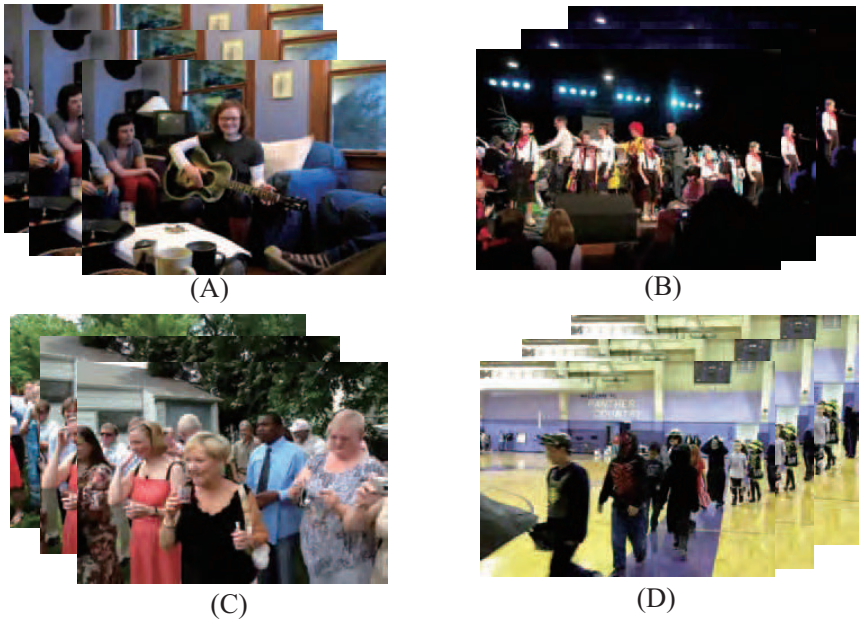


Figure 3.5: Examples of surprising videos: (A) birthday party with instrumental music, (B) music performance with costumes, (C) wedding ceremony with drinking glasses, (D) an indoor parade.

3.5.7 Further Evaluations

3.5.7.1 M2LATM improves robustness to label noise.

An important challenge for learning from real-world user data, or AMT annotations, is dealing with label-noise. We expect our model to deal better with label noise in the user-defined attributes, because it can additionally leverage automatically discovered latent attributes for a more robust overall representation. To simulate this, we repeated the previous multi-task and zero/N-shot learning experiments, but randomly flipped 50% of attribute bits on 50% of the training videos (so 25% wrong annotations). M2LATM is more robust than SVM-UD (Fig. 3.6 red vs blue), sometimes dramatically so. For example, when subjected to label noise, the multi-task classification performance of SVM-UD drops by 8% (vs only 3% for M2LATM) and actually performs worse than Direct.

3.5.7.2 User-defined and latent attributes should be learned jointly.

The M2LATM model has three complementary types of topics that define the semi-latent attribute space. An advantage of our model is to learn these jointly. To quantify this, we also learn them separately by training a batch of SVM classifiers (for UD topics), a constrained topic model (just CC topics), and an unsupervised topic model (GF topics). We compare performance using the

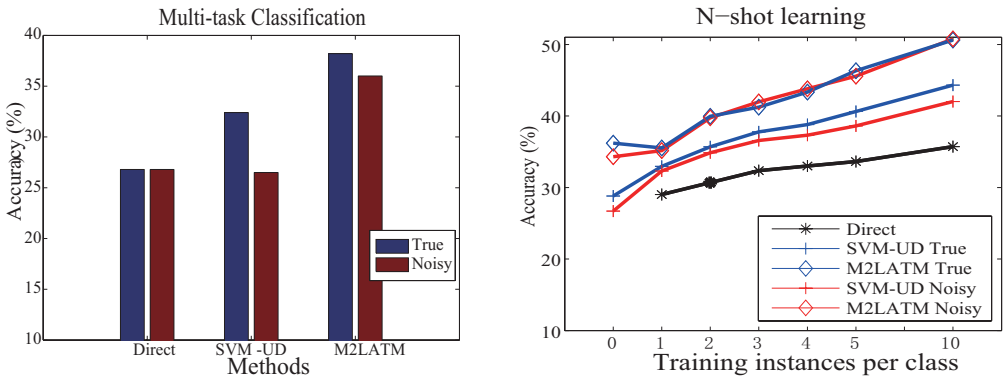


Figure 3.6: Robustness to attribute label-noise in multi-task classification and zero/N-shot learning.

	1-shot		5-shot		10-shot	
	Ind.	Joint	Ind.	Joint	Ind.	Joint
R/7	35.0	**38.8**	42.0	**48.0**	48.0	**52.2**
A/69	**38.8**	38.6	49.4	**52.1**	54.6	**56.1**

Table 3.5: Independent vs joint learning of semi-latent attributes. N-shot transfer. (4v4 classes, chance = 25%) .

concatenated output of the individual models vs the output of the jointly model in N-shot transfer learning. The results (Table 3.5) show that joint learning is always similar or significantly better than independent learning, so joint learning of latent attributes is indeed important to ensure they learn complementary aspects to UD attributes.

3.5.7.3 Significance of using SVM posteriors as user-defined attributes.

We use M2LATM to jointly learn UD, CC and GF attributes in a single generative model, with the aim of ensuring that latent attributes are complementary to user-defined attributes. However, as discussed in Section 3.5.3, we ultimately use the SVM posteriors in place of the UD topics because, being discriminatively trained strong classifiers, they perform slightly better. However, this is not a significant factor in our model's performance: across all the experiments, the margin of using SVM attribute classifiers over topic posteriors is $[-3\% \sim 4\%]$.

3.5.8 Analysis of Discovered Latent Attributes

3.5.8.1 *The setting of the number of latent attributes*

In each experiment, we fixed the attribute number to a round value greater than the total number of UD attributes in the exhaustive ontology. (69→100 for USAA, 85→150 for AwA). If too few are used, the model cannot learn anything beyond the UD ontology, if too many are used, the model can be over-fitting. The values used were set a priori and not tuned.

In general this free parameter could be eliminated by optimising it using cross-validation(CV)[3] or by developing a non-parametric Bayesian variant of our model to automatically determine the number of topics. The performance is not particularly sensitive to the number of topics. To illustrate point, we show some results for CCV performance while varying the number of topics in Figure 3.7:

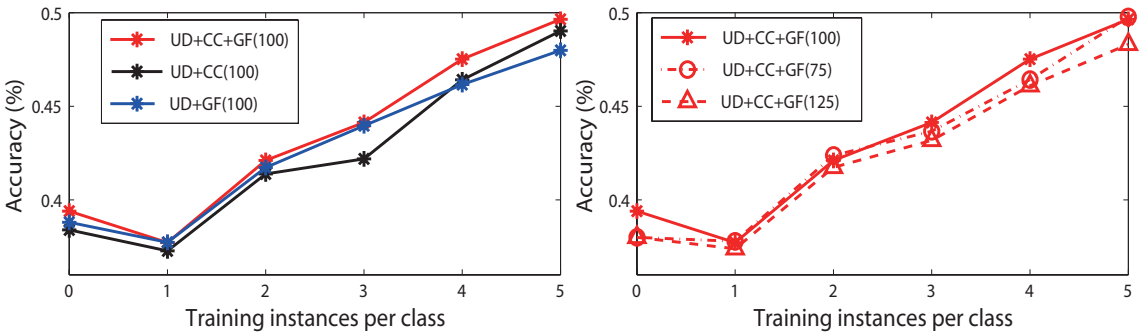


Figure 3.7: Zero and N-shot classification accuracy for USAA dataset. Left: Varying which type of latent attributes are included. Right: Varying total number of topics used.

3.5.8.2 *Latent attributes can discover user-defined attributes from a withheld ontology, as well as novel attributes outside the full ontology.*

In this section we investigate what is learned by latent attributes: can they discover UD attributes not provided in the ontology, and do they discover anything outside of the full UD ontology? Firstly, we define the distance between learned attributes i and j as the normalized correlation between their multinominal parameters $D(i,j) = \beta_i^T \beta_j / (\|\beta_i\| \, \|\beta_j\|)$. Fig. 3.8 shows the sorted similarity matrix between attributes for M2LATM learned in a conventional A/69 and semi-latent R/7 attribute setting. The diagonal structure shows that latent attributes have largely discovered many of the semantic UD attributes of interest to users. The uncorrelated strip to the right represents latent attributes in the R/7 model which have discovered aspects of the data not covered by

[3]However CV may be not reliable in the sparse data domain we are investigating

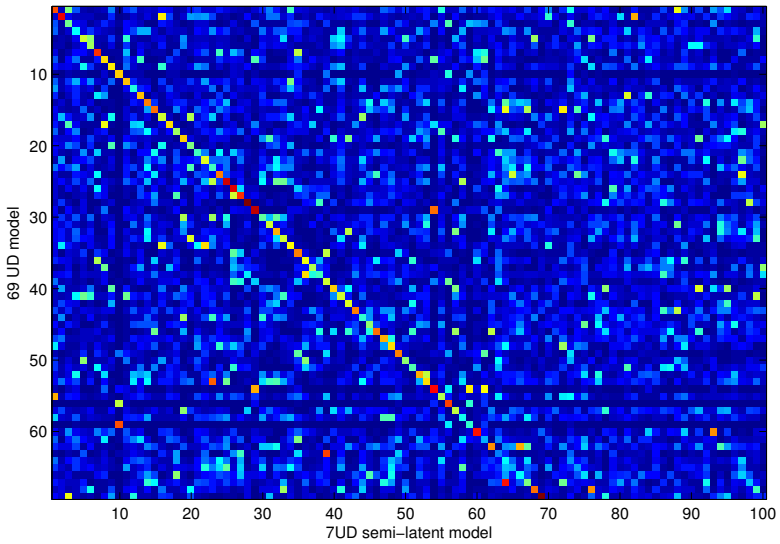


Figure 3.8: Similarity between user-defined and latent attributes.

the UD attributes.

To visualize an attribute, we select its top-N most likely words (from β), and then plot occurrences of these words on videos with high probability for this attribute (γ). Fig. 3.9 (top row) shows an example of static appearance (SIFT) attributes *bride* and *cake*. The high degree of overlapping between red circles and red crosses indicates that the re-discovered latent attributes match the withheld UD attributes well. Examples of STIP attributes *blow candle*, and *dancing* are shown in Fig. 3.9 (second row). For auditory attributes, we show *birthday song* and *speech* in Fig. 3.9 (third row). In this case, we plot the time-series of the attribute weight for the corresponding UD attribute and the latent attribute which rediscovered it along with ground-truth for when the particular sound was audible. All of these latent attributes were GF type, except *birthday song*, which was CC – being uniquely selective for birthday-party class.

Finally, to further illustrate the value of latent attributes, we visualized some latent attributes with no similarity to any existing UD attribute (i.e., those on the right strip of Fig. 3.8). This revealed new attributes which we had not included in our ontology despite intending it to be exhaustive. Fig. 3.9 (bottom row) shows two examples: (i) a *horizontal line* attribute, which the model learns is informative for classes with stages and fences such as concerts and performances; and (ii) a *tree* attribute, which the model learns is informative for typically outdoor or situations such as wedding receptions and parades. These results support our motivating point that manual ontologies are almost certainly *incomplete*, and benefit from being complemented with a set of latent attributes.

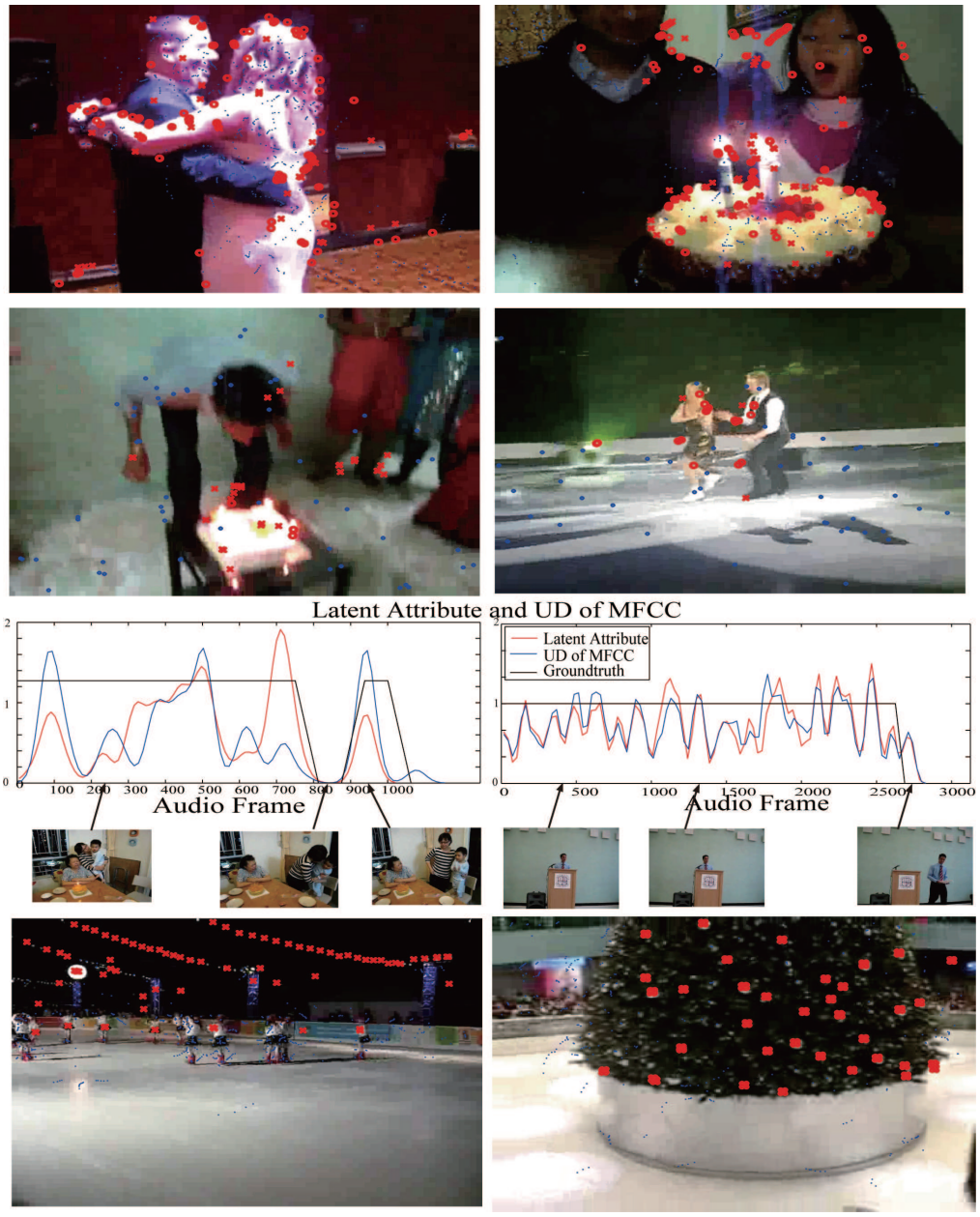


Figure 3.9: Visualization of user-defined (circles) and corresponding latent attributes (crosses). Red circles illustrate representative words from the UD attribute (A/69); red crosses illustrate the words from the corresponding latent attribute which discovered these concepts when withheld (R/7). Blue dots illustrate interest-points not related to attributes concerned.

	1-shot			5-shot			10-shot		
Condition	**Direct**	**SVM-UD**	**M2LATM**	**Direct**	**SVM-UD**	**M2LATM**	**Direct**	**SVM-UD**	**M2LATM**
N/0	16.4	-	**19.2**	21.5	-	**30.5**	23.6	-	**35.9**
R/9	16.4	25.1	**27.1**	21.5	32.6	**35.6**	23.6	36.4	**39.0**
R/42	16.4	**30.7**	28.3	21.5	**42.5**	**42.5**	23.6	45.0	**45.7**
A/85	16.4	**31.9**	28.5	21.5	**43.4**	38.0	23.6	**46.8**	43.0

Table 3.6: N-shot classification performance for AwA dataset (40v10 classes, chance = 10%).

3.5.9 Computational Scalability

In Section 3.4.2, we introduced a new sparse vocabulary-domain representation of our inference algorithm. To contrast the improved scalability of this representation (Eq. (3.9)) vs. the standard word-domain approach (Eq. (3.4), also used by [YA10, WBL09, NWFF08, FHXG12]), we recorded the matlab computation time for 10 instance multi-task learning on the USAA data. Our model required 30 minutes versus to 5 hours for the conventional approach. This margin grows with the video length and density of features, so this is an important contribution for scalability.

3.5.10 Experiments on Animals with Attributes (AwA)

Our model is not specific to videos/social activities. We also study the well known AwA dataset, (see [LNH09] for full details). AwA dataset defines 50 classes of animals, and 85 associated attributes (such as *furry*, and *has claws*). There are 30475 images with at least 92 examples of each class. We use the same six BoW features from [LNH09]. In contrast to USAA dataset, each class has a distinct deterministic definition in terms of attributes. For M2LATM, we keep the complexity fixed at 150 topics: with 1CC attribute per class, up to 85 user-defined attributes, and the others are GF latent attributes. There are six different kind of features extracted to describe the AwA images.

Table 3.6 shows N-shot learning results for AwA, with the attributes learned from all instances of 40 classes, and the target task learned from $1-10$ instances of the held out ten classes (same condition as [LNH09]). The same general results hold: M2LATM performs comparably or better than the others in most cases. Notably, although SVM-UD slightly outperforms M2LATM with the exhaustive A/85 condition (due to M2LATM's larger number of dimensions over fitting slightly), the use of latent attributes enables M2LATM to outperform SVM-UD in the most

		No Attrib Prior		Attrib Prior	
	[YA10]/[MSN11]/[KKTH12]	DAP	M2LATM	DAP	M2LATM
R/9	-	26.3	26.9	27.8	**29.2**
R/42	-	34.4	38.2	36.0	**39.7**
A/85	33.0/33.0/32.7	37.0	39.2	39.2	**41.3**

Table 3.7: Zero-shot classification performance (%) for AwA (40v10 classes, chance = 10%).

relevant and challenging cases of few UD attributes.

The ZSL results are shown in Table 3.7. Here, because the AwA attributes are deterministic, we were able to implement and apply DAP zero-shot learning precisely as described in [LNH09]. Different from [LNH09], we found that attribute priors provided a noticeable improvement in performance, so we show results with and without priors. In general, M2LATM outperforms DAP across the range of ontology completeness. For context, we also show the $\approx 33\%$ figure reported by several recent ZSL studies [YA10], [MSN11] and [KKTH12], although these conditions may not be exactly comparable to ours. This highlights the fact that our approach outperforms very recent methods with as few as half of the available attributes (R/42).

3.6 Summary

In this chapter a new framework is developed for multimedia understanding and focused on bridging the semantic and cross-modal gaps via an attribute-learning approach. In particular we focus on understanding videos of social group activities, which are particularly challenging and topical examples of this task because of their multi-modal content and complex and unstructured nature relative to the density of annotations. A solution to this problem would have huge application potential, e.g., content-based recognition and indexing, and hence content-based search, retrieval, filtering and recommendation of multimedia.

To solve this problem, we introduce the concept of semi-latent attribute space, expressing user-defined and latent attributes in a unified framework, and propose a novel scalable probabilistic topic model for learning multi-modal semi-latent attributes, which dramatically reduces requirements for an exhaustive accurate attribute ontology and expensive annotation effort. In experiments, we show that our framework is able to exploit latent attributes to outperform contemporary approaches for addressing a variety of realistic sparse multimedia data learning tasks

including: multi-task learning, learning with label noise, N-shot transfer learning and importantly zero-shot learning.

We address the limitations of previous studies including reliance on an exhaustive manual specification of the attribute space, ignoring or simplistically dealing with multi-modal content, and the unrealistic requirement of noiseless annotation of attributes. In particular, we are able to:

1. flexibly learn a full semantic-attribute space whether exhaustively defined, completely unavailable, available in a small subspace (i.e., present but sparse), or available but with noisy examples;

2. improve multi-task and N-shot learning by leveraging latent attributes;

3. go beyond existing zero-shot learning approaches (which only use user-defined attributes) by also exploiting latent attributes;

4. explicitly leverage attributes in conjunction with multi-modal data to improve cross-media understanding, enabling new tasks such as explicitly learning which modalities particular attributes appear in;

5. make our topic model applicable to large multimedia data by expressing it in a significantly more scalable way than previous studies – invariant to the length of the input video and density of the features.

Chapter 4

Transductive Multi-view Embedding

It is estimated that human can distinguish 30,000 basic object classes [Bie87] and many more subordinate ones (e.g. breeds of dogs). To recognise such high number of classes, humans have the ability to "learning to learn" and transfer knowledge from known classes to unknown ones. Inspired by this ability and to minimise the necessary labelled training examples for conventional supervised classifiers, researchers build the recognition models that are capable of classifying novel classes with no training example via attribute learning models. As introduced in Chapter 2, most attribute learning models learn a projection from a low-level feature space to the semantic space by the auxiliary dataset and apply such projection without adaptation to the target dataset. The knowledge is thus transferred from known classes to novel unknown classes with no or only a few labels.

Nevertheless there are three inherent problems that exist in previous attribute learning models. The more details of these problems have been discussed in Chapter 1.2.2, 1.2.3 and 1.2.4 respectively. We briefly summarise them here to give an overview in this Chapter.

Projection domain-shift problems: Since the known (auxiliary) and unknown (target) data have different and potentially unrelated classes, the underlying data distributions of the classes differ, so do the 'ideal' projection functions between the low-level feature space and the semantic spaces. Using the projection functions learned from the auxiliary dataset/domain without any adaptation to the target dataset/domain causes such an unknown shift/bias.

Prototype sparsity problems: For each target class, we only have a single prototype which is

insufficient to fully represent what that class looks like.

Inability to combine multiple semantic representations: Besides attribute representation, we can have multiple semantic representations (e.g. continuous word vectors [SGS+13]). Each representation (or semantic 'view') may contain complementary information – useful for distinguishing different classes in different ways. However, the state-of-the-art attribute learning algorithms are inable to explore multiple intermediate semantic representations.

Particularly, in the DAP model, different components of the model are affected by different problems and their negative effects aggregate and degrade the performance of zero-shot learning.

This chapter presents a transductive multi-view embedding framework to solve these three problems simultaneously. The transductive setting means using the unlabelled test data to improve generalisation accuracy. In our framework, each unlabelled target class instance is represented by multiple views: its low-level feature view and its (biased) projections in multiple semantic spaces (visual attribute space and word space in this work). To rectify the projection domain shift between auxiliary and target datasets, we introduce a multi-view semantic space alignment process to correlate different semantic views and the low-level feature view by projecting them onto a common latent embedding space learned using multi-view Canonical Correlation Analysis (CCA). The intuition is that when the biased target data projections (semantic representations) are correlated/aligned with their (unbiased) low-level feature representations, the bias/projection domain shift is alleviated. Furthermore, after exploiting the complementarity of different low-level feature and semantic views synergistically in the common embedding space, different target classes become more compact and more separable, making the subsequent zero-shot recognition a much easier task.

Even with the proposed transductive multi-view embedding framework, the prototype sparsity problem remains – instead of one prototype per class, a handful are now available depending on how many views are embedded, which are still sparse. Our solution is to pose this as a semi-supervised learning problem: prototypes in each view are treated as labelled 'instances', and we exploit the manifold structure of the unlabelled data distribution in each view in the embedding space via label propagation on a graph. To this end, we introduce a novel transductive multi-view hypergraph label propagation (TMV-HLP) algorithm for recognition. The core in our TMV-HLP algorithm is a new *distributed representation* of graph structure termed heterogeneous hypergraph which allows us to exploit the complementarity of different semantic and low-level feature

views, as well as the manifold structure of the target data to compensate for the impoverished supervision available from the sparse prototypes. Zero-shot learning is then performed by semi-supervised label propagation from the prototypes to the target data points within and across the graphs. The whole framework is illustrated in Figure 4.1.

By combining our transductive embedding framework and the TMV-HLP zero-shot recognition algorithm, our approach generalises seamlessly when none (zero-shot), or few (N-shot) samples of the target classes are available. Uniquely it can also synergistically exploit zero + N-shot (i.e., both prototypes and labelled samples) learning. Furthermore, the proposed method enables a number of novel cross-view annotation tasks.

The main content of this Chapter has been previously published/submitted in

1. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. "Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation" European Conference on Computer Vision (ECCV) 2014;

2. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gon"*Transductive Multi-view Zero-Shot Learning*" minor revision of IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)

3. Yanwei Fu, Yongxing Yang, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. "Transductive Multi-label Zero- shot Learning" British Machine Vision Conference (BMVC) 2014;

4. Yanwei Fu, Yongxing Yang, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. "Transductive Multi-calss and Multi-label Zero-shot Learning" ECCV 2014 workshop on Parts and Attribute;

4.1 Problem Setup of Transductive Multi-view Embedding Framework

We have c_S source/auxiliary classes with n_S instances $S = \{X_S, Y_S^i, \mathbf{z}_S\}$ and c_T target classes $T = \left\{X_T, Y_T^i, \mathbf{z}_T\right\}$ with n_T instances. X indicates the $t-$dimensional low-level feature of all instances; so $X_S \subseteq R^{n_s \times t}$ and $X_T \subseteq R^{n_T \times t}$. $\mathbf{z}_S$ and $\mathbf{z}_T$ are the auxiliary and target class label vectors. We assume the auxiliary and target classes are disjoint: $\mathbf{z}_S \cap \mathbf{z}_T = \varnothing$. We have I different types of intermediate semantic representations; Y_S^i and Y_T^i represent the ith type of m_i dimensional semantic representation for the auxiliary and target datasets respectively; so $Y_S^i \subseteq R^{n_S \times m_i}$ and

$Y_T^i \subseteq R^{n_T \times m_i}$. Note that for the auxiliary dataset, Y_S^i is given as each data point is labelled. But for the target dataset, Y_T^i is missing, and its prediction $\hat{Y}_T^i$ from X_T is used instead. As we will see later, this is obtained using a projection function learned from the auxiliary dataset. The problem of zero-shot learning is to estimate $\mathbf{z}_T$ given X_T and $\hat{Y}_T^i$.

4.2 Learning a Transductive Multi-View Embedding Space

Without any labelled data for the target classes, external knowledge is needed to represent what each target class looks like, in the form of class prototypes. Specifically, each target class c has a pre-defined class-level semantic prototype $\mathbf{y}_c^i$ in each semantic view i. In this Chapter, we consider two types of intermediate semantic representation (i.e. $I = 2$) – attributes and word vectors, which represent two distinct and complementary sources of information. We use $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{V}$ to denote the low-level feature, attribute and word vector spaces respectively. The attribute space $\mathcal{A}$ is typically manually defined using a standard ontology. For the word vector space $\mathcal{V}$, we employ the state-of-the-art skip-gram neural network model [MCCD13] trained on all English Wikipedia articles. To 13 Feb. 2014, it includes 2.9 billion words from a 4.33 million-words vocabulary (single and bi/tri-gram words). Using this learned model, we can project the textual name of any class into the $\mathcal{V}$ space to get its word vector representation. Unlike semantic attributes, it is a 'free' semantic representation in that this process does not need any human annotation. We next address how to project low-level features into these two spaces.

4.2.1 Learning the Projections of Semantic Spaces.

Mapping images and videos into a semantic space i requires a projection function $f^i : \mathcal{X} \rightarrow \mathcal{Y}^i$. This is typically realised by classifiers [LNH09] or regressors [SGS+13]. Using the auxiliary set S, we train support vector classifiers $f^{\mathcal{A}}(\cdot)$ and support vector regressors $f^{\mathcal{V}}(\cdot)$ for each dimension of the attribute and word vectors respectively[1]. Then the target class instances X_T have the semantic projections: $\hat{Y}_T^{\mathcal{A}} = f^{\mathcal{A}}(X_T)$ and $\hat{Y}_T^{\mathcal{V}} = f^{\mathcal{V}}(X_T)$. However, these predicted intermediate semantics have the projection domain shift problem as explained in Chapter 1. To solve this, we learn a transductive multi-view semantic embedding space to align the semantic projections with the low-level features of target data.

[1]Note that methods for learning projection functions for all dimensions jointly exist (e.g. [FCS+13]) and can be adopted in our framework.

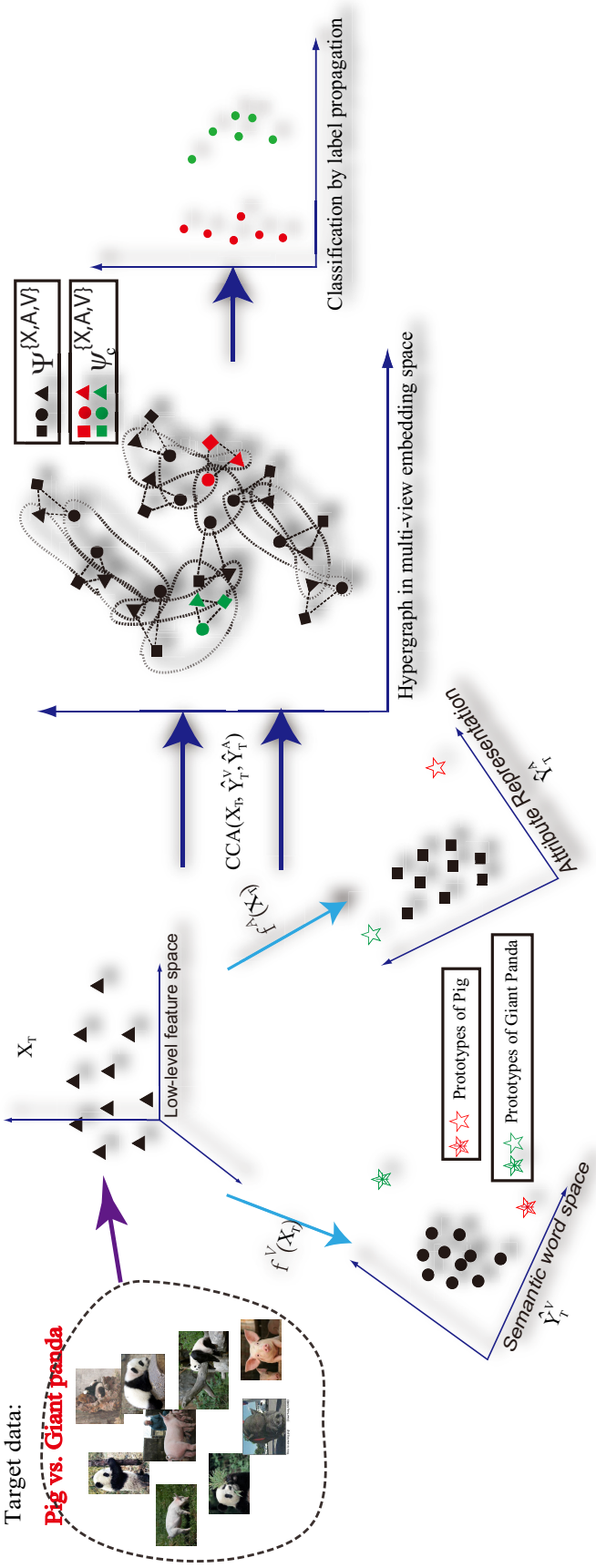


Figure 4.1: The pipeline of TMV-HLP framework.

4.2.2 Transductive Multi-view Embedding.

We introduce a multi-view semantic alignment (i.e. transductive multi-view embedding) process to correlate target instances in different (biased) semantic view projections with their low-level feature view. This process alleviates the projection domain shift problem, as well as providing a common space in which heterogeneous views can be directly compared, and their complementarity exploited then.

To learn an embedding space capable of rectifying the domain shift, we employ multi-view Canonical Correlation Analysis (CCA) for n_V views, with the target data representation in view i denoted as Φ^i, a $n_T \times m_i$ matrix. Specifically, in this work we project three views of each target class instance $f^{\mathcal{A}}(X_T)$, $f^{\mathcal{V}}(X_T)$ and X_T (i.e. $n_V = I + 1 = 3$) into a shared embedding space. The three projection functions W^i are learned by

$$\begin{array}{rcll} \min & \sum_{i,j=1}^{n_V} & Trace(W^i \Sigma_{ij} W^j) & \\ = & \sum_{i,j=1}^{n_V} & \| \Phi^i W^i - \Phi^j W^j \|_F^2 & \\ \text{s.t.} & \left[W^i\right]^T \Sigma_{ii} W^i = I & \left[\mathbf{w}_k^i\right]^T \Sigma_{ij} \mathbf{w}_l^j = 0 & \\ i \neq j, k \neq l & i,j = 1,\cdots,n_V & k,l = 1,\cdots,n_T & \end{array} \tag{4.1}$$

where W^i is the projection matrix which maps the view Φ^i (a n_T row matrix) into the embedding space and $\mathbf{w}_k^i$ is the kth column of W^i. Σ_{ij} is the covariance matrix between Φ^i and Φ^j. The optimisation problem above is multiconvex as long as Σ_{ii} are non-singular. The local optimum can be easily found by iteratively maximising over each W^i given the current values of the other coefficients as detailed in [HSST04].

The dimensionality of the embedding space is the sum of that of Φ^i, i.e. $\sum_{i=1}^{n_V} m_i$ – there is thus likely feature redundancy. Since the importance of each dimension is reflected by its corresponding eigenvalue [HSST04, GKIL13], we use the eigenvalues to weight the dimensions and define a *weighted embedding space* Γ:

$$\Psi^i = \Phi^i W^i \left[D^i\right]^\lambda = \Phi^i W^i \tilde{D}^i, \tag{4.2}$$

where D^i is a diagonal matrix with its diagonal elements set to the eigenvalues of each dimension in the embedding space, λ is a power weight of D^i and empirically set to 4 [GKIL13], and Ψ^i is the final representation of the target data from view i in Γ. We index the $n_V = 3$ views as $i \in \{\mathcal{X}, \mathcal{V}, \mathcal{A}\}$ for notational convenience. The same formulation can be used if more views are available.

4.2.3 Similarity in the Embedding Space.

The choice of similarity metric is important for high-dimensional embedding spaces [GKIL13]. In particular, extensive evidence in text analysis and information retrieval have shown that high-dimensional embedding vectors are naturally directional and using cosine similarity provides significant robustness against noise [BDGS05, GKIL13, HSST04]. Therefore for the subsequent recognition and annotation tasks, we compute cosine distance in Γ by l_2 normalisation: normalising any vector $\boldsymbol{\psi}_k^i$ (the kth row of Ψ^i to unit length (i.e. $\| \boldsymbol{\psi}_k^i \|_2 = 1$). Thus cosine similarity is given by the inner product of any two vectors in Γ. Finally, equipped with a weighted and normalised embedding space Γ, any two vectors can be directly compared no matter whether the original view is $\mathcal{X}$, $\mathcal{A}$ or $\mathcal{V}$.

4.3 Recognition by Multi-view Hypergraph Label Propagation (TMV-HLP)

After alleviating the projection domain shift problem by multi-view embedding, we next introduce a unified framework [FYH+14a] – TMV-HLP to fuse multiple views and transductively exploit the manifold structure of the unlabelled target data to perform zero-shot, as well as N-shot learning if sparse labelled samples for the target classes are available.

Each target class is defined by a single semantic prototype in each semantic view, which can be a binary attribute vector, or the class name represented as a word vector in the word space. Such class-level prototypes are effectively the expected mean for the distribution of this class in semantic space, since the projection function f^i aims at mapping each instance to be near to its class prototype in each semantic view. Formally, we assume a target class c has a prototype $\mathbf{y}_c^i$ in each semantic view for zero-shot, and/or a few labelled instances for N-shot classification. To exploit the learned embedding space Γ for recognition, we project three views of each unlabelled target instance $f^{\mathcal{A}}(X_T)$, $f^{\mathcal{V}}(X_T)$ and X_T as well as the target class prototypes into Γ [2]. The prototypes $\mathbf{y}_c^i$ for views $i \in \{\mathcal{A}, \mathcal{V}\}$ are projected as $\boldsymbol{\psi}_c^i = \mathbf{y}_c^i W^i \tilde{D}^i$. So we have $\boldsymbol{\psi}_c^{\mathcal{A}}$ and $\boldsymbol{\psi}_c^{\mathcal{V}}$ for the attribute and word vector prototypes of each target class c in Γ. In the absence of a prototype for the (non-semantic) low-level feature view $\mathcal{X}$, we synthesise it as $\boldsymbol{\psi}_c^{\mathcal{X}} = (\boldsymbol{\psi}_c^{\mathcal{A}} + \boldsymbol{\psi}_c^{\mathcal{V}})/2$.

Most or all of the target instances are unlabelled, so we leverage graph-based semi-supervised learning to exploit the manifold structure of the unlabelled data in each view transductively for classification. This differs from the conventional approaches such as direct attribute prediction

[2] Before being projected into Γ, the prototypes are updated by one-step self-training as in [FHXG13].

(DAP) [LNH13] which essentially assume that the data distribution for each target class is Gaussian or multinomial. Since the directional high-dimensional data in our embedding space lies on a unit-sphere manifold [GY14], their assumptions are invalid and the proposed graph-based label propagation algorithm is more appropriate. However, since our embedding space contains multiple projections of the target data, it is hard to define a single graph that synergistically exploits the manifold structure of all views. We therefore consider the generalised graph-based framework.

4.3.1 The Overview of Multi-view Hypergraph Label Propagation (TMV-HLP)

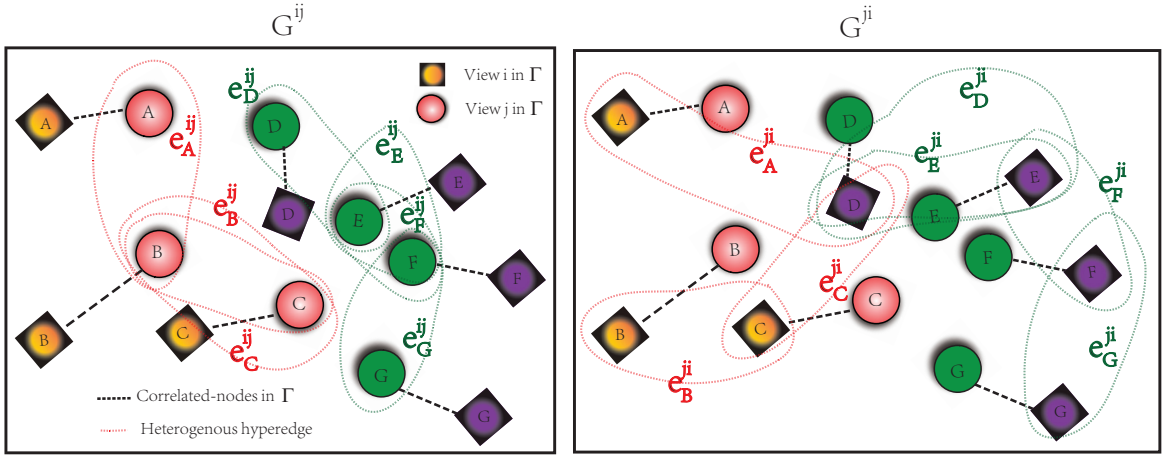


Figure 4.2: An example of constructing heterogeneous hypergraphs. Suppose in the embedding space, we have 14 nodes belonging to 7 data points A, B, C, D, E, F and G of two views – view i (rectangle) and view j (circle). Data points A, B, C and D, E, F, G belong to two different classes – red and green respectively. The multi-view semantic embedding maximises the correlations (connected by black dash lines) between the two views of the same node. Two hypergraphs are shown ($\mathcal{G}^{ij}$ at the left and $\mathcal{G}^{ij}$ at the right) with the heterogeneous hyperedges drawn with red/green dash ovals for the nodes of red/green classes. Each hyperedge consists of two most similar nodes to the query node.

We introduce the transductive multi-view hypergraph label propagation framework. Specifically, thanks to the shared embedding space Γ, we construct heterogeneous hypergraphs across views to combine/align the different manifold structures so as to further enhance the robustness and exploit the complementarity of different views. These graphs become comparable and can be connected by a Bayesian prior weight estimated from data. Given the constructed graphs, the

initial label information from the prototypes (zero-shot) and/or the few labelled target data points (N-shot) is then propagated to the unlabelled data by random walk on the graphs.

Before fully developing these two frameworks, we define the pairwise node similarity in the embedding space Γ.

4.3.2 Pairwise Node Similarity

The key idea behind such graph-based methods is to group similar data points, represented as vertices/nodes on a graph, into edges/hyperedges. With such edges/hyperedges, the pairwise similarity between two data points is measured as the similarity between the two edges/hyperedges that they belong to, instead of that between the two nodes only. Before that, pairwise similarity between two graph nodes needs to be defined. In our embedding space Γ, each data point in each view defines a node, and the similarity between any pair of nodes is:

$$\omega(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j) = \exp(\frac{< \boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j >^2}{\varpi}) \tag{4.3}$$

where $< \boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j >^2$ is the square of inner product between the ith and jth projections of nodes k and l with a bandwidth parameter ϖ. Most previous work [RES13, ZB07] sets ϖ by cross validation which is not possible for zero-shot learning. Inspired by [Lam09], a simple strategy for setting ϖ is adopted: $\varpi \approx \underset{k,l=1,\cdots,n}{\text{median}} < \boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j >^2$ in order to have roughly the same number of similar and dissimilar sample pairs. This makes the edge weights from different pairs of nodes more comparable. Note that Eq (4.3) defines the pairwise similarity between any two nodes within the same view ($i = j$) or across different views ($i \neq j$).

The whole pipeline of TMV-HLP is illustrated in Figure 4.1. We next discuss how to construct the multi-view heterogenous hypergraph and the corresponding label propagation by random walk.

4.3.3 Heterogeneous Hyperedges

Given the multi-view projections of the target data, we aim to construct a set of across-views heterogeneous hypergraphs $\mathcal{G}^c = \left\{\mathcal{G}^{ij} \mid i, j \in \{\mathcal{X}, \mathcal{V}, \mathcal{A}\}, i \neq j\right\}$. Within the set, a cross-view heterogeneous hypergraph for views i and j (in that order) is denoted as $\mathcal{G}^{ij} = \left\{\Psi^i, E^{ij}, \Omega^{ij}\right\}$ where Ψ^i is the node set in view i, E^{ij} is the hyperedge set and Ω^{ij} is the pairwise node similarity set for the hyperedges. Given each node in i, denoted as $\boldsymbol{\psi}_k^i$, a hyperedge $e^{ij}_{\boldsymbol{\psi}_k^i}$ is constructed by

searching for similar nodes in view j. We thus have

$$E^{ij} = \left\{ e^{ij}_{\boldsymbol{\psi}^i_k} \mid i \neq j, k = 1, \cdots n_T + c_T \right\} \tag{4.4}$$

where each hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$ includes the nodes in view j that are the nearest to node $\boldsymbol{\psi}^i_k$ in view i and the similarity set

$$\Omega^{ij} = \left\{ \Delta^{ij}_{\boldsymbol{\psi}^i_k} = \left\{ \omega\left(\boldsymbol{\psi}^i_k, \boldsymbol{\psi}^j_l\right) \right\} \mid i \neq j, \boldsymbol{\psi}^j_l \in e^{ij}_{\boldsymbol{\psi}^i_k} \, k = 1, \cdots n_T + c_T \right\} \tag{4.5}$$

where $\omega\left(\boldsymbol{\psi}^i_k, \boldsymbol{\psi}^j_l\right)$ is computed using Eq (4.3). Since the hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$ intrinsically groups all nodes in view j that are most similar to node $\boldsymbol{\psi}^i_k$ in view i, we call $\boldsymbol{\psi}^i_k$ the query node for hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$. Similarly, $\mathcal{G}^{ji}$ can be constructed by using nodes in j to query nodes in i. Therefore given three views, we have six across view/heterogeneous hypergraphs. Figure 4.2 illustrates two heterogeneous hypergrahs constructed from two views. Interestingly, our way of defining hyperedges naturally corresponds to the star expansion [SJY08] where the query node (i.e. $\boldsymbol{\psi}^i_k$) is introduced to connect each node in the hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$.

4.3.4 Similarity Strength Between Hyperedge and Query Node

For each hyperedge, we measure its similarity strength with its query node which will be used later to compute similarity between two hyperedges. Specifically, we use the weight $\delta^{ij}_{\boldsymbol{\psi}^i_k}$ to indicate the similarity strength of nodes connected within each heterogeneous hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$. Thus, we define $\delta^{ij}_{\boldsymbol{\psi}^i_k}$ based on the mean similarity of the set $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$ for the hyperedge

$$\delta^{ij}_{\boldsymbol{\psi}^i_k} = \frac{1}{\mid e^{ij}_{\boldsymbol{\psi}^i_k} \mid} \sum_{\omega\left(\boldsymbol{\psi}^i_k, \boldsymbol{\psi}^j_l\right) \in \Delta^{ij}_{\boldsymbol{\psi}^i_k}, \boldsymbol{\psi}^j_l \in e^{ij}_{\boldsymbol{\psi}^i_k}} \omega\left(\boldsymbol{\psi}^i_k, \boldsymbol{\psi}^j_l\right), \tag{4.6}$$

where $\mid e^{ij}_{\boldsymbol{\psi}^i_k} \mid$ is the cardinality of hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$.

Due to the multi-view embedding step and the way of setting ϖ in Eq (4.3), the similarity sets $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$ and $\Delta^{ij}_{\boldsymbol{\psi}^i_l}$ can generally be directly compared. Nevertheless, to make subsequent computation more robust, we use the following normalisation of the similarity sets: (a) we assume $\forall \Delta^{ij}_{\boldsymbol{\psi}^i_k} \in \Omega^{ij}$ and $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$ should follow Gaussian distribution. Thus, we enforce zero-score normalisation to $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$; (b) We further assume that the retrieved similarity set Ω^{ij} between all the queried nodes $\boldsymbol{\psi}^i_k$ ($l = 1, \cdots n_T$) from view i and $\boldsymbol{\psi}^j_l$ should also follow Gaussian distribution. So we again enforce Gaussian distribution to the pairwise similarities between $\boldsymbol{\psi}^j_l$ and all query nodes from view i by zero-score normalisation; (c) We select the first k highest values from $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$ as new similarity set $\bar{\Delta}^{ij}_{\boldsymbol{\psi}^i_k}$ for hyperedge $e^{ij}_{\boldsymbol{\psi}^i_k}$. $\bar{\Delta}^{ij}_{\boldsymbol{\psi}^i_k}$ is then used in Eq (4.6) in place of $\Delta^{ij}_{\boldsymbol{\psi}^i_k}$. This normalisation steps aim to compute a more robust similarity between each pair of hyperedges.

4.3.5 Pairwise Hyperedge Similarity

For each hyperedge there is an associated incidence matrix $H^{ij} = \left(h\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right)\right)_{(n_T+c_T)\times|E^{ij}|}$ where

$$h\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right) = \begin{cases} 1 & if\ \boldsymbol{\psi}_l^j \in e^{ij}_{\boldsymbol{\psi}_k^i} \\ 0 & otherwise \end{cases} \tag{4.7}$$

To take into consideration the similarity strength between hyperedge and query node, we extend the binary valued hyperedge incidence matrix H^{ij} to soft-assigned incidence matrix $SH^{ij} = \left(sh\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right)\right)_{(n_T+c_T)\times|E^{ij}|}$ as follows

$$sh\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right) = \delta^{ij}_{\boldsymbol{\psi}_k^i} \cdot \omega\left(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j\right) \cdot h\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right) \tag{4.8}$$

This soft-assigned incidence matrix is the product of three components: (1) the weight $\delta_{\boldsymbol{\psi}_k^i}$ for hyperedge $e^{ij}_{\boldsymbol{\psi}_k^i}$; (2) the pairwise similarity computed using queried node $\boldsymbol{\psi}_k^i$; (3) the binary valued hyperedge incidence matrix element $h\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right)$. To make the values of SH^{ij} comparable among the different heterogeneous views, we apply l_2 normalisation to the soft-assigned incidence matrix values for all node incident to each hyperedge.

Now for each heterogeneous hypergraph, we can finally define the pairwise similarity between any two nodes or hyperedges. Specifically for $\mathcal{G}^{ij}$, the similarity between the oth and lth nodes is

$$\omega_c^{ij}\left(\boldsymbol{\psi}_o^j, \boldsymbol{\psi}_l^j\right) = \sum_{e^{ij}_{\boldsymbol{\psi}_k^i} \in E^{ij}} sh\left(\boldsymbol{\psi}_o^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right) \cdot sh\left(\boldsymbol{\psi}_l^j, e^{ij}_{\boldsymbol{\psi}_k^i}\right). \tag{4.9}$$

With the pairwise hyperedge similarity, one can now create the hypergraphs. In principle, one could create a hypergraph where all hyperedges are exhaustively connected; however, that is too costly for the subsequent label propagation task, so we use a k-nearest-neighbour (kNN) graph [Zhu07]. In this work we set $k = 30$, which still can be varied from $10 \sim 50$ with little effects in our experiments.

4.3.6 The Advantages of Heterogeneous Hypergraphs

We argue that the pairwise similarity of heterogeneous hypergraph is a distributed representation [Ben09] in Eq (4.9). To explain it, we can use star extension [SJY08] to extend a hypergraph into a traditional 2-graph. For each hyperedge $e^{ij}_{\boldsymbol{\psi}_k^i}$, the query node $\boldsymbol{\psi}_k^i$ is used to compute the pairwise similarity $\Delta^{ij}_{\boldsymbol{\psi}_k^i}$ of all the nodes in view j. Each hyperedge can thus define a hyper-plane by categorising the nodes in view j into two groups: strong and weak similarity group regarding

to query node $\boldsymbol{\psi}_k^i$. In other words, the hyperedge set E^{ij} is multi-clustering with linearly separated regions (by each hyperplane) per classes. Since the final pairwise similarity in Eq (4.9) can be represented by a set of similarity weights computed by hyperedge, and such weights are not mutually exclusive and are statistically independent, we consider the heterogeneous hypergraph a distributed representation. The advantage of having a distributed representation has been studied by Watts and Strogatz [WS98, Wat04] which shows that such a representation gives rise to better convergence rate and better clustering abilities. In contrast, the homogeneous hypergraphs adopted by previous work [HLZM10, FGZ$^+$10, HYLC13] does not have this property which makes them less robust against noise. In addition, fusing different views in the early stage of graph construction potentially can lead to better exploitation of the complementarity of different views. The advantages over homogeneous hypergraphs are validated by our experiments. However, it is worth pointing out that (1) The reason we can query nodes across views to construct heterogeneous hypergraph is because we have projected all views in the same embedding space in the first place. (2) Hypergraphs typically gain robustness at the cost of losing discriminative power – it essentially blurs the boundary of different clusters/classes by taking average over hyperedges. A typical solution to this is to fuse hypergraphs with the conventional 2-graphs [FGZ$^+$10, HYLC13], which is adopted in this work as well.

4.3.7 Label Propagation by Random Walk

Now we have two types of graphs: heterogeneous hypergraphs $\mathcal{G}^c = \left\{\mathcal{G}^{ij}\right\}$ and 2-graphs $\mathcal{G}^p = \left\{\mathcal{G}^i\right\}$. Given three views ($n_V = 3$), we thus have nine graphs in total (six hypergraphs and three 2-graphs). To propagate label information from prototypes/labelled nodes to other unlabelled nodes, a classic strategy is random walk [ZB07]. We next define a random walk process within and across graphs. A natural random walk on $\mathcal{G} = \{\mathcal{G}^p; \mathcal{G}^c\}$ for two nodes k and l has the following transition probability,

$$p\left(k \to l\right) = \sum_{i \in \{\mathcal{X},\mathcal{V},\mathcal{A}\}} p\left(k \to l \mid \mathcal{G}^i\right) \cdot p\left(\mathcal{G}^i \mid k\right) + \sum_{i,j \in \{\mathcal{X},\mathcal{V},\mathcal{A}\}, i \neq j} p\left(k \to l \mid \mathcal{G}^{ij}\right) \cdot p\left(\mathcal{G}^{ij} \mid k\right) \tag{4.10}$$

where

$$p\left(k \to l \mid \mathcal{G}^i\right) = \frac{\omega_p^i(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^i)}{\sum_o \omega_p^i(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_o^i)}, \tag{4.11}$$

and

$$p\left(k \to l \mid \mathcal{G}^{ij}\right) = \frac{\omega_c^{ij}(\boldsymbol{\psi}_k^j, \boldsymbol{\psi}_l^j)}{\sum_o \omega_c^{ij}(\boldsymbol{\psi}_k^j, \boldsymbol{\psi}_o^j)}$$

and then the posterior probability to choose graph $\mathcal{G}^i$ at projection/node $\boldsymbol{\psi}_k^i$ will be:

$$p(\mathcal{G}^i|k) = \frac{\pi(k|\mathcal{G}^i)p(\mathcal{G}^i)}{\sum_i \pi(k|\mathcal{G}^i)p(\mathcal{G}^i) + + \sum_{ij} \pi(k|\mathcal{G}^{ij})p(\mathcal{G}^{ij})} \tag{4.12}$$

$$p(\mathcal{G}^{ij}|k) = \frac{\pi(k|\mathcal{G}^{ij})p(\mathcal{G}^{ij})}{\sum_i \pi(k|\mathcal{G}^i)p(\mathcal{G}^i) + \sum_{ij} \pi(k|\mathcal{G}^{ij})p(\mathcal{G}^{ij})} \tag{4.13}$$

where $p(\mathcal{G}^i)$ and $p(\mathcal{G}^{ij})$ are the prior probability of graphs $\mathcal{G}^i$ and $\mathcal{G}^{ij}$ in the random walk. This probability expresses prior expectation about the informativeness of each graph. The same Bayesian model averaging [FHX+14a] can be used here to estimate these prior probabilities. However, the computational cost is combinatorially increased with the number of views; and it turns out the prior is not critical to the results of our framework. Therefore, uniform prior is used in our experiments.

The stationary probabilities for node k in $\mathcal{G}^i$ and $\mathcal{G}^{ij}$ are

$$\pi(k|\mathcal{G}^i) = \frac{\sum_l \omega_p^i(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^i)}{\sum_o \sum_l \omega_p^i(\boldsymbol{\psi}_k^i, \boldsymbol{\psi}_o^i)} \tag{4.14}$$

$$\pi(k|\mathcal{G}^{ij}) = \frac{\sum_l \omega_c^{ij}(\boldsymbol{\psi}_k^j, \boldsymbol{\psi}_l^j)}{\sum_k \sum_o \omega_c^{ij}(\boldsymbol{\psi}_k^j, \boldsymbol{\psi}_o^j)} \tag{4.15}$$

Finally, the stationary probability across the multi-view hypergraph is computed as:

$$\pi(k) = \sum_{i \in \{\mathcal{X}, \mathcal{V}, \mathcal{A}\}} \pi(k|\mathcal{G}^i) \cdot p(\mathcal{G}^i) + \tag{4.16}$$

$$\sum_{i,j \in \{\mathcal{X}, \mathcal{V}, \mathcal{A}\}, i \neq j} \pi(k|\mathcal{G}^{ij}) \cdot p(\mathcal{G}^{ij}) \tag{4.17}$$

Given the defined graphs and random walk process, we can derive our label propagation algorithm (TMV-HLP). Let P denote the transition probability matrix defined by Eq (4.10) and Π the diagonal matrix with the elements $\pi(k)$ computed by Eq (4.16). The Laplacian matrix $\mathcal{L}$ combines information of different views and is defined as: $\mathcal{L} = \Pi - \frac{\Pi P + P^T \Pi}{2}$. The label matrix Z for labelled N-shot data or zero-shot prototypes is defined as:

$$Z(q_k, c) = \begin{cases} 1 & q_k \in class\, c \\ -1 & q_k \notin class\, c \\ 0 & unknown \end{cases} \tag{4.18}$$

Given the label matrix Z and Laplacian $\mathcal{L}$, label propagation on multiple graphs has the closed-form solution [ZB07] : $\hat{Z} = \eta(\eta\Pi + \mathcal{L})^{-1}\Pi Z$ where η is a regularisation parameter[3]. Note that in our framework, both labelled target class instances and prototypes are modelled as graph nodes. Thus the difference between zero-shot and N-shot learning lies only on the initial labelled instances: Zero-shot learning has the prototypes as labelled nodes; N-shot has instances as labelled nodes; and a new condition exploiting both prototypes and N-shot together is possible. This unified recognition framework thus applies when either or both of prototypes and labelled instances are available. The computational cost of our TMV-HLP is $\mathcal{O}\left(k \cdot (c_T + n_T)^2 \cdot n_V^2\right)$, where k is the number of nearest neighbours in the kNN graphs, and n_V is the number of views.

4.4 Annotation and Beyond

Our multi-view embedding space Γ bridges the semantic gap between low-level features $\mathcal{X}$ and semantic representations $\mathcal{A}$ and $\mathcal{V}$. Leveraging this cross-view mapping, annotation [HGX11a, WG07, GKIL13] can be improved and applied in novel ways. We consider three annotation tasks here:

4.4.1 Instance Level Annotation

Given a new instance u, we can describe/annotate it by predicting its attributes. The conventional solution is directly applying $\hat{\mathbf{y}}_u^{\mathcal{A}} = f^{\mathcal{A}}(\mathbf{x}_u)$ for test data $\mathbf{x}_u$ [FEHF09, GKIL13]. However, as analysed before, this suffers from the projection domain shift problem. To alleviate this problem, our multi-view embedding space aligns the semantic attribute projections with the low-level features of each unlabelled instance in the target domain. Such an alignment can thus be used for image annotations of the target domain. Thus, with our framework, we can now infer attributes for any test instance via the learned embedding space Γ as $\hat{\mathbf{y}}_u^{\mathcal{A}} = \mathbf{x}_u W^{\mathcal{X}} \tilde{D}^{\mathcal{X}} \left[W^{\mathcal{A}} \tilde{D}^{\mathcal{A}}\right]^{-1}$.

4.4.2 Zero-shot Class Description

From a broader machine intelligence perspective, one might be interested to ask what are the attributes of an unseen class, based solely on the class name. By virtue of our multi-view embedding space, *zero-shot class description* can be performed to infer the semantic attribute description of a novel class. This *zero-shot class description* task could be useful, for example, to hypothesise the zero-shot attribute prototype of a class instead of defining it by experts [LNH09]

[3] It can be varied from $1-10$ with little effects in our experiments

or ontology [FHXG13]. Our transductive embedding space enables this task by connecting semantic word space (i.e. naming) and discriminative attribute space (i.e. describing). Given the prototype $\mathbf{y}_c^{\mathcal{V}}$ from the name of a novel class c, we compute $\hat{\mathbf{y}}_c^{\mathcal{A}} = \mathbf{y}_c^{\mathcal{V}} W^{\mathcal{V}} \tilde{D}^{\mathcal{V}} \left[W^{\mathcal{A}} \tilde{D}^{\mathcal{A}}\right]^{-1}$ to generate the class-level attribute description.

4.4.3 Zero Attribute Learning

This task is the inverse of the previous task – to infer the name of class given a set of attributes. It could be useful, for example, to validate or assess a proposed zero-shot attribute prototype, or to provide an automated semantic-property based index into a dictionary or database. To our knowledge, this is the first attempt for evaluating the quality of a class attribute prototype because no previous work has directly and systematically linked linguistic knowledge space with visual attribute space. Specifically given an attribute prototype $\mathbf{y}_c^{\mathcal{A}}$, we can use $\hat{\mathbf{y}}_c^{\mathcal{V}} = \hat{\mathbf{y}}_c^{\mathcal{A}} W^{\mathcal{A}} \tilde{D}^{\mathcal{A}} \left[W^{\mathcal{V}} \tilde{D}^{\mathcal{V}}\right]^{-1}$ to name the corresponding class and perform retrieval on dictionary words in $\mathcal{V}$ using $\hat{\mathbf{y}}_c^{\mathcal{V}}$.

4.5 Experiments

4.5.1 Datasets And Settings.

We evaluate our framework on three widely used image/video attribute datasets: Animal with Attribute (AwA), Unstructured Social Activity Attribute (USAA), and Caltech-UCSD-Birds (CUB). For detailed of these datasets, please refer to Chapter 2.1.6.

AwA [LNH09] consists of 50 classes of animals (30475 images) and 85 associated class-level attributes. It has a standard source/target split for zero-shot learning with 10 classes and 6180 images held out as the target dataset. We use the same 'hand-crafted' low-level features (RGB colour histograms, SIFT, rgSIFT, PHOG, SURF and local self-similarity histograms) released with the dataset (denoted as $\mathcal{H}$); and the same multi-kernel learning (MKL) attribute classifier from [LNH09].

USAA [FHXG13] is a video dataset with 69 instance-level attributes for 8 classes of complex (unstructured) social group activity videos from YouTube. Each class has around 100 training and test videos respectively. USAA provides the instance-level attributes since there are significant intra-class variations. We use the thresholded mean of instances from each class to define a binary attribute prototype as in Chapter 3 as well as [FHXG13]. The

same setting in Chapter 3 is adopted: 4 classes as source and 4 classes as target data. We use exactly the same SIFT, MFCC and STIP low-level features for USAA as in [FHXG13].

CUB-200-2011 [WBW+11] contains 11,788 images of 200 bird classes. This is a more challenging dataset than AwA – it is designed for fine-grained recognition and has more classes but fewer images. Each class is annotated with 312 binary attributes derived from the bird species ontology. We use 150 classes as auxiliary data, holding out 50 as target data. We extract 128 dimensional SIFT and colour histogram descriptors from regular grid of multi-scale and aggregate them into image-level feature Fisher Vectors ($\mathcal{F}$) by using 256 Gaussians, also as in [APHS13]. Colour histogram and PHOG features are also used to extract the global color and texture information from each image. Due to the recent progress on deep learning based representations, we also extract the Overfeat ($\mathcal{O}$) [SEZ+14][4] from AwA and CUB as an alternative to $\mathcal{H}$ and $\mathcal{F}$ respectively. In addition, the Decaf ($\mathcal{D}$) [DJV+14] [5] feature is also considered for AwA.

We report absolute classification accuracy on USAA and mean accuracy for AwA and CUB for direct comparison to published results. The word vector space is trained by the skip model [MCCD13] with 1000 dimensions.

4.5.2 Recognition by Zero-shot Learning

Comparisons with state-of-the-art. We compare our method – TMV-HLP with most recent state-of-the-art models that report results or can be reimplemented by us on the three datasets in Table 4.1. They cover a wide range of approaches on utilising semantic intermediate representation for zero-shot learning. They can be roughly categorised according to the semantic representation(s) used: DAP and IAP ([LNH09], [LNH13]), M2LATM in Chapter 3 (as well as [FHXG13]), ALE [APHS13], [RES13] and [WJ13] use attributes only; HLE/AHLE [APHS13] and Mo/Ma/O/D [RSS+10] use both attributes and linguistic knowledge bases (same as us); [YCF+13] uses attribute and some additional human manual annotation. Note that our linguistic knowledge base representation is in the form of word vectors, which does not incur additional manual annotation. Our method also does not exploit data-driven attributes such as M2LATM in Chapter 3 (as well as [FHXG13]) and Mo/Ma/O/D [RSS+10].

[4] We use the trained model of Overfeat in [SEZ+14].

[5] Provided at `http://attributes.kyb.tuebingen.mpg.de/`.

Approach	AwA ($\mathcal{H}$ [LNH09])	AwA ($\mathcal{O}$)	AwA ($\mathcal{O},\mathcal{D}$)	USAA	CUB ($\mathcal{O}$)	CUB ($\mathcal{F}$)
DAP	40.5([LNH09]) / 41.4([LNH13]) / 38.4*	51.0*	57.1*	33.2([FHXG13]) / 35.2*	26.2*	9.1*
IAP	27.8([LNH09]) / 42.2([LNH13])	–	–	–	–	–
M2LATM (Chapter 3)	41.3	–	–	41.9	–	–
ALE/HLE/AHLE [APHS13]	37.4/39.0/43.5	–	–	–	–	18.0
Mo/Ma/O/D [RSS$^+$10]	27.0 / 23.6 / 33.0 / 35.7	–	–	–	–	–
PST [RES13]	42.7	54.1*	62.9*	36.2*	38.3*	13.2*
[WJ13]	43.4	–	–	–	–	–
[YCF$^+$13]	48.3**	–	–	–	–	–
TMV-HLP	**49.0**	**73.5**	**80.5**	**50.4**	**47.9**	**19.5**

Table 4.1: Comparison with the state-of-the-art on zero-shot learning on AwA, USAA and CUB. Features $\mathcal{H}$, $\mathcal{O}$ and $\mathcal{F}$ represent hand-crafted, OverFeat and Fisher Vector respectively. Mo, Ma, O and D represent the highest results in the mined object class-attribute associations, mined attributes, objectiveness as attributes and direct similarity methods used in [RSS$^+$10] respectively. '–': no result reported. *: our implementation. **: requires additional human annotations.

Let us first look at the results on the most widely used AwA. Apart from the results obtained with the standard hand-crafted feature ($\mathcal{H}$), we consider the more powerful Overfeat deep feature ($\mathcal{O}$), and a combination of Overfeat and Decaf $(\mathcal{O},\mathcal{D})$[6]. Table 4.1 shows that (1) with the same experimental settings and the same feature ($\mathcal{H}$), our TMV-HLP outperforms the best result reported so far (48.3%) in [YCF$^+$13] which requires additional human annotation to relabel the similarities between auxiliary and target classes. In contrast, our method uses no additional human annotation. (2) With the more powerful Overfeat feature, our method achieves 73.5% zero-shot recognition accuracy. Even more remarkably, when both the Overfeat and Decaf features are used in our framework, the result (see the AwA $(\mathcal{O},\mathcal{D})$ column) is 80.5%. Even with only 10 classes, this is an extremely good result given that we do not have any labelled samples from target classes. Note that this result is not solely due to the feature strength, as the margin between the conventional DAP and our TMV-HLP is much bigger indicating that our TMV-HLP plays a critical role in achieving this result. We will have more in-depth analysis on this result later. (3) Our method is also superior to the AHLE method in [APHS13] which also uses two semantic spaces: attribute and WordNet hierarchy. Different from our embedding framework,

[6]With these two low-level feature views, there are six views in total in the embedding space.

AHLE simply concatenates the two spaces. (4) Our method also outperforms the other alternatives of either mining other semantic knowledge bases (Mo/Ma/O/D [RSS+10]) or exploring data-driven attributes (M2LATM in Chapter 3). (5) Among all compared methods, PST [RES13] is the only one except ours that performs label propagation based transductive learning. It yields better results than DAP in all the experiments which essentially does nearest neighbour in the semantic space. TMV-HLP consistently beats PST in all the results shown in Table 4.1 thanks to our multi-view embedding.

Table 4.1 also shows that on two very different datasets: USAA video activity, and CUB fine-grained, our TMV-HLP significantly outperforms the state-of-the-art alternatives. In particular, on the more challenging CUB, 47.9% accuracy is achieved on 50 classes (chance level 2%) using the Overfeat feature. Considering the fine-grained nature and the number of classes, this is even more impressive than the 80.5% result on AwA. It is also noted that again the advantage of our TMV-HLP is even clearer when using the more powerful deep learning feature than the conventional Fisher Vector feature ($\mathcal{F}$).

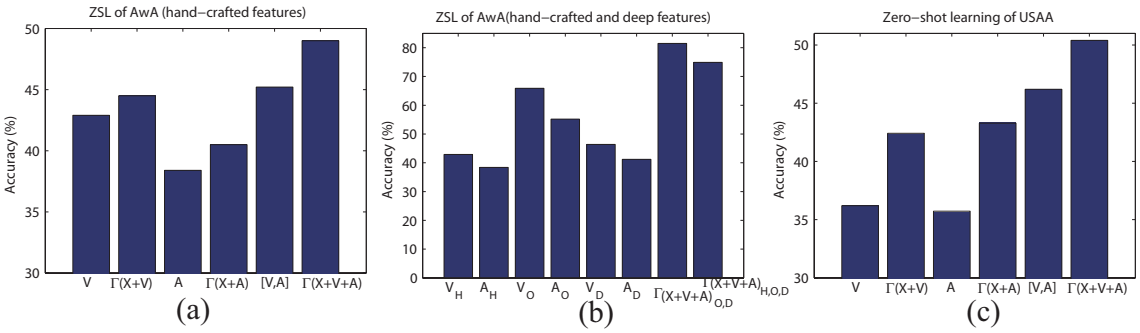


Figure 4.3: Effectiveness of transductive multi-view embedding. (a) zero-shot learning on AwA using only hand-crafted features; (b) zero-shot learning on AwA using hand-crafted and deep features together; (c) zero-shot learning on USAA. $[\mathcal{V},\mathcal{A}]$ indicates the concatenation of semantic word and attribute space vectors. $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$ mean using low-level+semantic word spaces and low-level+attribute spaces respectively to learn the embedding. $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ indicates using all 3 views to learn the embedding.

4.5.3 Transductive multi-view embedding helps

To validate the contribution of our transductive multi-view embedding space we split up different views with and without embedding and the results are shown in Fig. 4.3. In Figs. 4.3(a) and (c), the hand-crafted feature $\mathcal{H}$ and Fisher Vector $\mathcal{F}$ are used for AwA and CUB respectively, and

we compare $\mathcal{V}$ vs. $\Gamma(\mathcal{X}+\mathcal{V})$, $\mathcal{A}$ vs. $\Gamma(\mathcal{X}+\mathcal{A})$ and $[\mathcal{V},\mathcal{A}]$ vs. $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ (see the caption of Figure. 4.3 for definitions). We use DAP for $\mathcal{A}$ and nearest neighbour for $\mathcal{V}$ and $[\mathcal{V},\mathcal{A}]$, because the prototypes of $\mathcal{V}$ are not binary vectors so DAP cannot be applied. We use TMV-HLP for $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$ respectively. We highlight the following observations: (1) After transductive embedding, $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$, $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$ outperform $[\mathcal{V},\mathcal{A}]$, $\mathcal{V}$ and $\mathcal{A}$ respectively. This means that the transductive embedding is helpful whichever semantic space is used in rectifying the projection domain shift problem by aligning the semantic views with low-level features. (2) The results of $[\mathcal{V},\mathcal{A}]$ are higher than those of $\mathcal{A}$ and $\mathcal{V}$ individually, showing that the two semantic views are indeed complementary even with simple feature level fusion. However, our TMV-HLP on all views $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ improves individual embeddings $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$.

4.5.3.1 Embedding deep learning feature views also helps and the more views the better

In Fig. 4.3(b) three different low-level features are considered for AwA: hand-crafted ($\mathcal{H}$), overfeat ($\mathcal{O}$) and decaf features ($\mathcal{D}$). The zero-shot learning results of each individual space are indicated as $\mathcal{V}_\mathcal{H}$, $\mathcal{A}_\mathcal{H}$, $\mathcal{V}_\mathcal{O}$, $\mathcal{A}_\mathcal{O}$, $\mathcal{V}_\mathcal{D}$, $\mathcal{V}_\mathcal{D}$ in Figure 4.3(b) and we observe that $\mathcal{V}_\mathcal{O} > \mathcal{V}_\mathcal{D} > \mathcal{V}_\mathcal{H}$ and $\mathcal{A}_\mathcal{O} > \mathcal{A}_\mathcal{D} > \mathcal{A}_\mathcal{H}$. That is Overfeat $>$ Decaf $>$ hand-crafted features. It is widely reported that deep features have better performance than 'hand-crafted' features on many computer vision benchmark datasets [SEZ+14, CSVZ14]. What is interesting to see here is that Overfeat $>$ Decaf since both are based on the same Convolutional Neural Network (CNN) model of [KSH12]. Apart from implementation details, one significant difference is that Decaf is pre-trained by ILSVRC2012 while Overfeat by ILSVRC2013 which contains more animal classes meaning better (more relevant) features can be learned. It is also worth pointing out that

1. with both Overfeat and Decaf features, the number of views to learn embedding space doubles from 3 to 6; and our results suggest that the more views, the better chance to solve the domain shift problem and the data become more separable as different views contain complementary information;

2. Figure 4.3(b) shows that when all 9 available views ($\mathcal{X}_\mathcal{H}$, $\mathcal{V}_\mathcal{H}$, $\mathcal{A}_\mathcal{H}$, $\mathcal{X}_\mathcal{D}$, $\mathcal{V}_\mathcal{D}$, $\mathcal{A}_\mathcal{D}$, $\mathcal{X}_\mathcal{O}$, $\mathcal{V}_\mathcal{O}$ and $\mathcal{A}_\mathcal{O}$) are used for embedding, the result is significantly better than those from from each individual view. Nevertheless, it is lower than that obtained by embedding views ($\mathcal{X}_\mathcal{D}$, $\mathcal{V}_\mathcal{D}$, $\mathcal{A}_\mathcal{D}$, $\mathcal{X}_\mathcal{O}$, $\mathcal{V}_\mathcal{O}$ and $\mathcal{A}_\mathcal{O}$). This suggests that view selection may be required when a large number of views are available for learning the embedding space. However this is

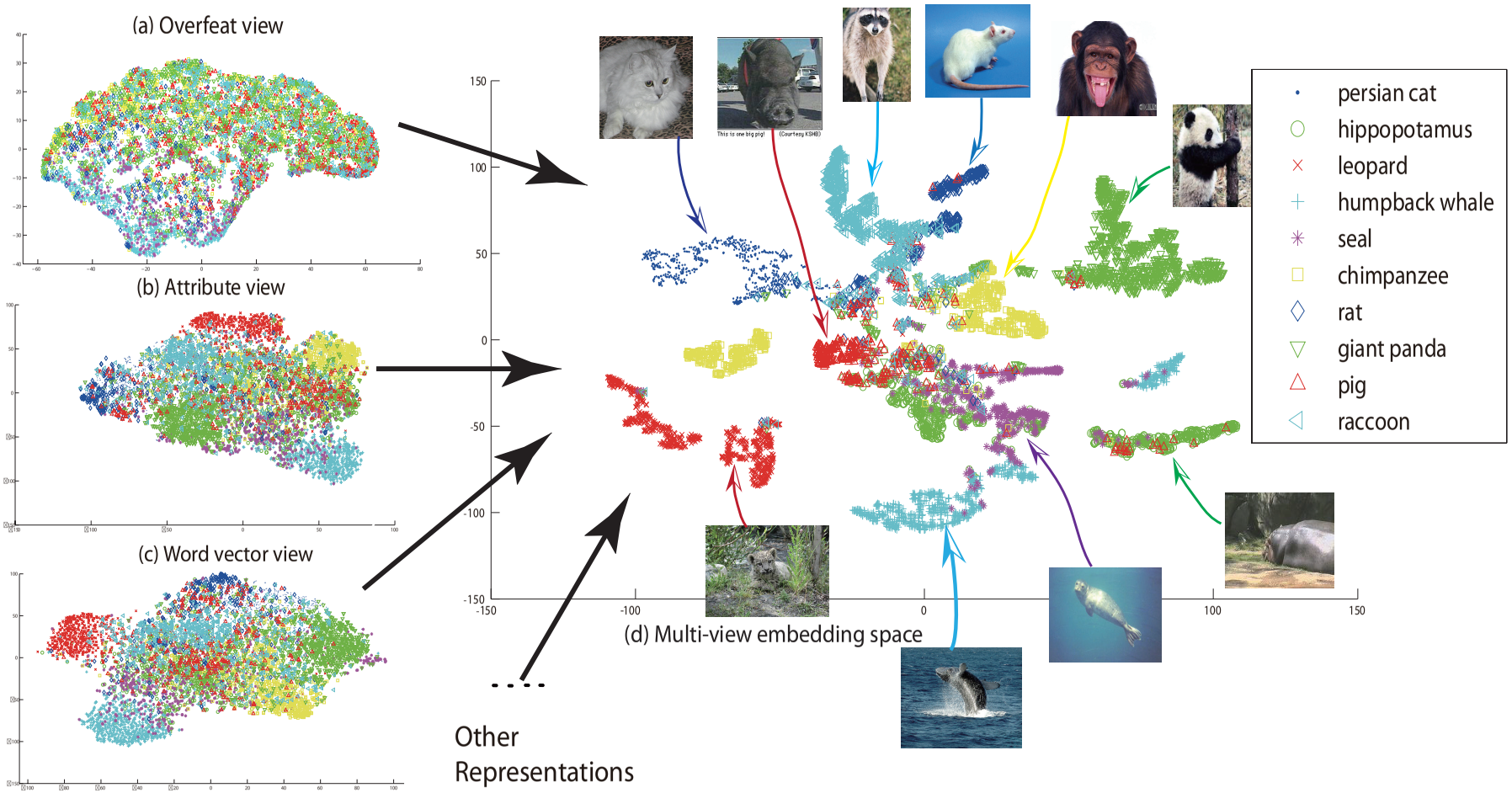


Figure 4.4: Visualisation of (a) Overfeat view ($\mathcal{X}_{\mathcal{O}}$), (b) attribute view ($\mathcal{A}_{\mathcal{O}}$), (c) word vector view ($\mathcal{V}_{\mathcal{O}}$), and (d) the pairwise graph generated by TMV-HLP ($\Gamma(\mathcal{X}+\mathcal{A}+\mathcal{V})_{\mathcal{O},\mathcal{D}}$) in multi-view embedding space . The unlabelled target classes are much more separable in (d).

non-trivial for zero-shot learning as there is no validation set.

4.5.3.2 Embedding makes different classes more separable

We employ t-SNE [vdMH08] to visualise the space $\mathcal{X}_{\mathcal{O}}$, $\mathcal{V}_{\mathcal{O}}$, $\mathcal{A}_{\mathcal{O}}$ and $\Gamma(\mathcal{X}+\mathcal{A}+\mathcal{V})_{\mathcal{O},\mathcal{D}}$ in Figure. 4.4. We use the same default parameters of t-SNE[7] for each individual space.

It shows that even in the powerful Overfeat view, the 10 target classes are heavily overlapped (Figure. 4.4(a)). It gets better in the semantic views (Figure. 4.4(b) and (c)). However, when all 6 views are embedded, all classes are clearly separable (Figure 4.4(d)).

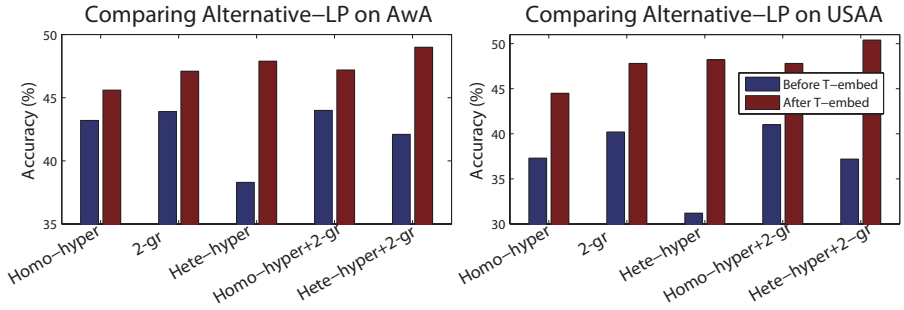


Figure 4.5: Comparing alternative label propagation methods. The methods differ in the graph models used (see text for details).

4.5.3.3 Heterogeneous hypergraph vs. other graphs

Apart from transductive multi-view embedding, our another major contribution is a novel label propagation method based on heterogeneous hypergraphs. To evaluate the effectiveness of our hypergraph label propagation, we compare with a number of alternative label propagation methods using other graph models. More specifically, within each view, two alternative graphs can be constructed: 2-graphs which are used in the classification on multiple graphs (C-MG) model [ZB07], and conventional homogeneous hypergraph formed in each single view [ZHS06, FGZ+10, LLS+13]. Since hypergraphs are typically combined with 2-graphs, we have 5 different multi-view graph models: *2-gr* (2-graph in each view), *Homo-hyper* (homogeneous hypergraph in each view), *Hete-hyper* (our heterogeneous hypergraph across views), *Homo-hyper+2-gr* (homogeneous hypergraph combined with 2-graph in each view), and *Hete-hyper+2-gr* (our heterogeneous hypergraph combined with 2-graph, as in our TMV-HLP). In our experiments, the same random walk label propagation algorithm is run on each graph in AwA and USAA before and after transductive embedding to compare these models.

[7]Similar visualisation schemes such as isomap, MDS and sammon mapping can also be used here; however, t-SNE is shown more robust than these methods in [vdMH08].

From the results in Figure 4.5, we observe that:

1. The graph model used in our TMV-HLP (*Hete-hyper+2-gr*) yields the best performance on both datasets.

2. All graph models benefit from the embedding. In particular, the performance of our heterogeneous hypergraph degrades drastically without embedding. This is expected because before embedding, nodes in different views are not aligned; so forming meaningful hyperedges across views is not possible.

3. Fusing hypergraphs with 2-graphs helps – as discussed above, one has the robustness and the other has the discriminative power, so it makes sense to combine the strengths of both.

4. After embedding, on its own, heterogeneous graphs are the best while homogeneous hypergraphs (*Homo-hyper*) are worse than 2-gr meaning the discriminative power by 2-graphs over-weighs the robustness of homogeneous hypergraphs.

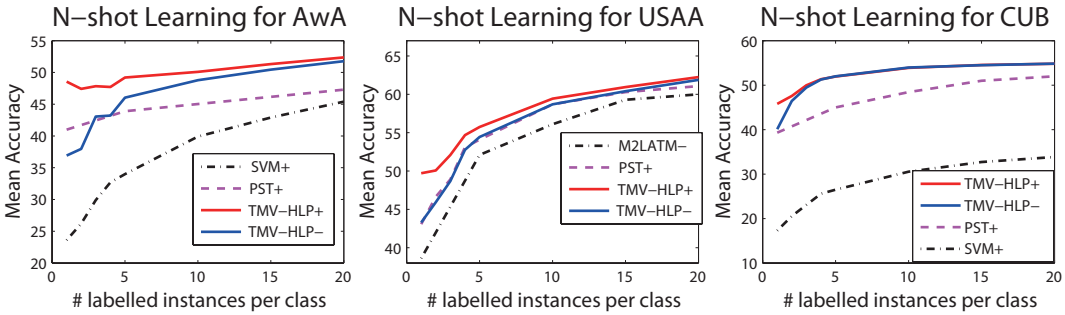


Figure 4.6: N-shot learning results.

4.5.3.4 *Qualitative results*

Figure 4.7 shows qualitative results for zero-shot learning on AwA in terms of top 5 most likely classes predicted for each image. TMV-HLP produces more reasonable ranked list of classes for each image.

4.5.3.5 *Running time*

Our TMV-HLP algorithm is computationally efficient. For example, our pipeline of using hand-crafted features on AwA dataset takes less than 30 minutes on a platform with six 2.66-GHz CPU cores for the zero-shot learning classification task (over 6,180 images). This includes the time for multi-view CCA embedding and label propagation using our heterogeneous hypergraphs.

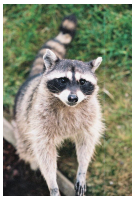
			
TMV-HLP	**giant panda**, leopard, seal, rat, raccoon	**raccoon**, seal, persian cat, leopard, chimpanzee	**rat**, persian cat, chimpanzee, seal, raccoon
PST	**giant panda**, seal, raccoon, rat, leopard	leopard, humpback whale, **raccoon**, chimpanzee, persian cat	chimpanzee, **rat**, persian cat, raccoon, seal
DAP	leopard, **giant panda**, raccoon, seal, chimpanzee	**raccoon**, chimpanzee, leopard, humpback whale, seal	leopard, giant panda, seal, hippopotamus, persian cat

Figure 4.7: Qualitative results for zero-shot learning on AwA.

4.5.4 N-Shot learning

N-shot learning experiments are carried out on the three datasets with the number of target class instances labelled (N) ranging from 0 (zero-shot) to 20. We also consider the situation [RES13] where both a few training examples *and* a zero-shot prototype may be available (denoted with suffix $+$), and contrast it to the conventional N-shot learning without the prototypes (denoted with suffix $-$). For comparison, PST+ is the method in [RES13] which uses prototypes for the initial label matrix. SVM+ and M2LATM- are the SVM and M2LATM methods used in [LNH13] and Chapter 3 respectively. For fair comparison, we modify the SVM- used in [LNH13] into SVM+. Note that our TMV-HLP can be used in both conditions but the PST method [RES13] only applies to the $+$ condition. All experiments are repeated for 10 rounds with the average results reported. Evaluation is done on the remaining unlabelled target data. From the results shown in Figure 4.6, it can be seen that:

(1) TMV-HLP+ always achieves the best performance, particularly given few training examples.

(2) The methods that explore transductive learning via label propagation (e.g., TMV-HLP+, TMV-HLP-, and PST+) are clearly superior to those that do not (e.g., SMV+ and M2LATM-).

(3) On AwA, PST+ outperforms TMV-HLP- with less than 3 instances per class. Be-

cause PST+ exploits the prototypes, this suggests that a single good prototype is more informative than a few labelled instances in N-shot learning. This also explains why sometimes the N-shot learning results of TMV-HLP+ are worse than its zero-shot learning results when only few training labels are observed (e.g. on AwA, the TMV-HLP+ accuracy goes down before going up when more labelled instances are added). Note that when more labelled instances are available, TMV-HLP- starts to outperform PST+, because it combines the different views of the training instances, and the strong effect of the prototypes is eventually outweighed.

4.5.5 Annotation And Beyond

In this section we evaluate our multi-view embedding space for the conventional and novel annotation tasks introduced in Sec. 4.4.

4.5.5.1 Instance annotation by attributes

To quantify the annotation performance, we predict attributes/annotations for each target class instance for USAA, which has the largest instance level attribute variations among the three datasets. We employ two standard measures: mean average precision (mAP) and F-measure (FM) between the estimated and true annotation list. Using our multi-view embedding space, our method (FM:0.341, mAP: 0.355) outperforms significantly the baseline of directly estimating $\mathbf{y}_u^{\mathcal{A}} = f^{\mathcal{A}}(\mathbf{x}_u)$ (FM:0.299, mAP: 0.267).

4.5.5.2 Zero-shot description

In this task, we explicitly infer the attributes corresponding to a specified novel class, given only the textual name of that class without seeing any visual samples. Table 4.3 illustrates this for AwA. Clearly most of the top/bottom 5 attributes predicted for each of the 10 target classes are meaningful (in the ideal case, all top 5 should be true positives and all bottom 5 true negatives). Quantitatively, given the top-5 attributes predicted for each class, a F-measure of 0.236 is obtained. In comparison, if we directly select the 5 nearest attribute name projection to the class name projection (prototype) in the word space, the F-measure value becomes 0.063, demonstrating the importance of learning the multi-view embedding space. In addition to providing a method to automatically – rather than manually – generate an attribute ontology, this task is interesting because even a human could find it very challenging (effectively a human has to list the attributes of a class which he has never seen or been explicitly taught about, but has only seen

mentioned in text).

4.5.5.3 Zero-attribute learning

In this task we attempt the reverse of the previous experiment: inferring a class name given a list of attributes. Table 4.2 illustrates this for USAA. Table 4.2(a) shows queries in USAA (note that class name is shown for brevity, but it is the attributes of those classes that are queried) and the top-4 ranked list of classes returned. The estimated class names of each attribute vector are reasonable – the top-4 words are either the class name or related to the class name. A baseline is to use the textual names of the attributes projected in the word space (using the sum of their word vectors) to search for the nearest class in the word space, instead of the embedding space. Table 4.2(a) shows that the predicted classes in this case are still reasonable, but significantly worse than querying via the embedding space. To quantify this we evaluate the average rank of the true name for each USAA class when queried by its attributes. For querying by embedding space, the average rank of the true class is an impressive 2.13 (out of 4.33M words with a chance-level rank of 2.17M), compared with the average rank of 110.24 by directly querying word space [MCCD13] with textual descriptions of the attributes. Table 4.2(b) shows an example of "incremental" query using the ontology definition of birthday party defined in Chapter 3. We first query the *wrapped presents* attribute only, followed by adding *small balloon* and all other attributes (*birthday songs* and *birthday caps*). The changing list of top ranked retrieved words intuitively reflects the expectation of the combinatorial meaning of the attributes.

(a)	Query via embedding space	Query attribute words in word space
graduation party	**party**, **graduation**, audience, caucus	cheering, proudly, dressed, wearing
music_performance	**music**, **performance**, musical, heavy metal	sing, singer, sang, dancing
wedding_ceremony	**wedding_ceremony**, wedding, glosses, stag	nun, christening, bridegroom, **wedding_ceremony**

(b) Attribute Query	Top Ranked Words
wrapped presents	music; performance; solo_performances; performing
+small balloon	wedding; wedding_reception; birthday_celebration; birthday
+All attributes	**birthday_party**; prom; wedding reception

Table 4.2: Zero-attribute learning on USAA. (a) Querying class names by attributes of classes. (b) An incrementally constructed attribute query for the birthday_party class. Bold indicates true positive words retrieved.

4.6 Summary

We identified the challenge of projection domain shift in zero-shot learning and presented a new framework to solve it by rectifying the biased projections in a multi-view embedding space. We also proposed a novel label-propagation algorithm TMV-HLP based on heterogeneous across-view hypergraphs. TMV-HLP synergistically exploit multiple intermediate semantic representations, as well as the manifold structure of unlabelled target data to improve recognition in a unified way for zero shot, N-shot and zero+N shot learning tasks. As a result we achieved state-of-the-art performance on the challenging AwA, CUB and USAA datasets. Finally, we demonstrated that our framework enables novel tasks of relating textual class names and their semantic attributes.

AwA		Attributes
pc	T-5	active, **furry, tail, paws, ground.**
	B-5	swims, hooves, long neck, horns, arctic
hp	T-5	**old world, strong, quadrupedal**, fast, **walks**
	B-5	red, plankton, skimmers, stripes, tunnels
lp	T-5	**old world, active, fast, quadrupedal, muscle**
	B-5	plankton, arctic, insects, hops, tunnels
hw	T-5	**fish**, **smart, fast, group, flippers**
	B-5	hops, grazer, tunnels, fields, plains
seal	T-5	**old world, smart**, **fast**, chew teeth, **strong**
	B-5	fly, insects, tree, hops, tunnels
cp	T-5	**fast, smart, chew teeth, active**, brown
	B-5	tunnels, hops, skimmers, fields, long neck
rat	T-5	**active, fast, furry, new world, paws**
	B-5	arctic, plankton, hooves, horns, long neck
gp	T-5	**quadrupedal**, active, **old world**, **walks**, **furry**
	B-5	tunnels, skimmers, long neck, blue, hops
pig	T-5	**quadrupedal**, **old world**, **ground**, furry, **chew teeth**
	B-5	desert, long neck, orange, blue, skimmers
rc	T-5	**fast**, **active**, **furry**, **quadrupedal**, **forest**
	B-5	long neck, desert, tusks, skimmers, blue

Table 4.3: Zero-shot description of the 10 AwA target classes. The embedding space is learned using 6 views ($\mathcal{X}_{\mathcal{D}}$, $\mathcal{V}_{\mathcal{D}}$, $\mathcal{A}_{\mathcal{D}}$, $\mathcal{X}_{\mathcal{O}}$, $\mathcal{V}_{\mathcal{O}}$ and $\mathcal{A}_{\mathcal{O}}$). The true positives are highlighted in bold. lp, pc, hp, hw, gp, rc and cp are short for leopard, Persian cat, hippopotamus, humpback whale, giant panda, raccoon, and chimpanzee respectively.

Chapter 5

Robust Learning of Relative Attributes

Attributes can be annotated at the instance-level or class-level. The instance-level attribute annotation is more desirable for the transfer learning problems, since more semantic information is labelled for each instance. Nevertheless, it is impossible to collect a large number of instance annotations if only in the laboratory environment. As a result, some crowdsourcing tools are employed for example, Amazon Mechanical Turk (AMT). Each instance is thus annotated by users to indicate the presence/absence of certain properties in the image or video. Nevertheless, such 'binary' attributes are not intrinsically versatile enough to express more 'specific' semantic meanings, for example, the relative information of any two instances.

From a much broader perspective, relative attribute is one special type of subjective visual properties; and this chapter actually studies the problems of robustly estimating subjective visual properties, which nevertheless encompass a variety of important applications. For example: estimating attractiveness [PG11b] from faces would interest social media or online dating websites; and estimating properties of consumer goods such as shininess of shoes [KPG12] improves customer experiences on online shopping websites. Recently, the problem of automatically predicting if people would find an image or video interesting has started to receive increasing attention [DOB11, GGR+13, JYF+13]. Interestingness prediction has a number of real-world applications. In particular, since the number of images and videos uploaded to the Internet is growing explosively, people are increasingly relying on image/video search engines or recommendation tools to select which ones to view. Given a query, ranking the retrieved data with relevancy to the query based on the predicted interestingness would improve the user satisfaction. Similarly

user stickiness can be increased if a media-sharing website such as YouTube can recommend videos that are both relevant and interesting. Other applications such as web advertising and video summarisation can also benefit. Subjective visual properties such as the above-mentioned ones are useful on their own. But they can also be used as an intermediate representation for other tasks such as visual recognition, e.g., different people can be recognised by how pale their skin complexions are and how chubby their faces look like [PG11b]. When used as a semantically meaningful representation, these subjective visual properties often are referred to as relative attributes [KPG12, PG11b, ZWX+13].

To capture more general semantic relationships and better understand complex multi-modal visual data, relative attributes were recently introduced as a richer semantic representation corresponding to the strengh of visual properties and used to describe relative information of any two instances. Such relative attributes are annotated by crowdsourcing tools e.g. AMT. The labelling process is more scalable and economic than the conventional laboratory annotation, but this crowdsourced labelling still suffers from sparsity and outliers problems, which are detailed in Chapter 1.2.5 and this chapter recaps below,

Sparsity: The number of pairwise comparisons required is much bigger than the number of data points because n instances defining a $\mathcal{O}(n^2)$ pairwise space. Consequently, even with crowdsourcing tools, the annotation remains be sparse, i.e. not all pairs are compared and each pair is only compared a few times.

Outliers: The crowd is not all trustworthy: it is well known that crowdsourced data are greatly affected by noise and outliers [CB13, WHG11, LHK13] which can be caused by a number of factors. Some workers may be lazy or malicious [KCS08], providing random or wrong annotations either carelessly or intentionally; some other outliers are unintentional human errors caused by the ambiguous nature of the data, thus are unavoidable regardless how good the attitudes of the workers are.

It is thus a challenge to robustly learn relative attributes from crowdsourced pairwise comparisons.

In this chapter, we propose a more principled way to identify annotation outliers by formulating the relative attribute prediction task as a unified robust learning to rank problem, tackling both the outlier detection and the prediction tasks jointly. Different from previous work which

replies on majority voting or statistical-based algorithms to prune the outliers and make the annotation less subjective and more reliable, our method operates globally, integrating all local pairwise comparisons together to minimise a cost that corresponds to global inconsistency of ranking order. This enables us to identify outliers that receive majority votes and yet cause large global ranking inconsistency and thus should be removed. Extensive experiments in this chapter on several benchmark datasets demonstrate that our new approach significantly outperforms state-of-the-art alternatives.

The main content of this Chapter has been previously published in

1. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Shaogang Gong and Yuan Yao. "Interestingness Prediction by Robust Learning to Rank" European Conference on Computer Vision (ECCV) 2014;

2. Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Shaogang Gong, Yizhou Wang and Yuan Yao."*Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels*"submitted to IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI);

5.1 A Unified Robust Learning to Rank (URLR) Framework

We aim to learn an relative attribute prediction model from a set of sparse and noisy pairwise comparisons, each comparison corresponding to a local ranking between a pair of images or videos. A Unified Robust Learning to Rank (URLR) framework is proposed for such purpose in this section.

5.1.1 Problem Setup

Suppose our training set has I data points/instances represented by a low-level feature matrix $\Phi = \left[\phi_i^T\right]_{i=1}^I \in R^{I\times d}$, where ϕ_i is a d-dimensional column feature vector for representing instance i. The annotations or data labels are represented as an annotation matrix Y. In particular, assume each pair of instances on average receive K votes by annotators. We will have $Y_{ij}^k = 1$ if the k-th vote indicates that instance i is more interesting than instance j, and $Y_{ji}^k = 1$ otherwise. The annotation matrix is then constructed as $Y_{ij} = \frac{1}{K}\sum_k Y_{ij}^k$. These pairwise comparisons can be naturally represented by a directed graph $G = (V,E)$ with node set $V = \{i\}_{i=1}^I$ and edge set $E = \{i \to j | Y_{ij} > 0\}$. That is, an edge $i \to j$ exists if $Y_{ij} > 0$.

Given the training data Φ and Y, there are two tasks: (1) removing the outliers in Y and (2) estimating a prediction function for the relative attribute values. In this Chapter a linear function is considered due to its low computational complexity, that is, given the low-level feature ϕ_x of a test instance x we use a linear function $f(x) = \beta^T \phi_x$ to predict its relative attribute values, where β is the coefficient weight vector of the low-level feature ϕ_x. All formulations can be easily updated to use a non-linear function.

Note that the annotation matrix Y is not symmetric – in an ideal case, one hopes that the votes received on each pair are unanimous, e.g. $Y_{ij} > 0$ and $Y_{ji} = 0$; but often there are disagreements, i.e. both $Y_{ij} > 0$ and $Y_{ji} > 0$. Assuming both cannot be true simultaneously, one of them will be an outlier. In this case, one is the majority and the other minority which will be pruned by the majority voting method. This is why majority voting is a local outlier detection method and requires as many votes per pair as possible to be effective (the wisdom of a crowd). Note that Y_{ij} and Y_{ji} indicate the contradictory voting and it is also possible that both Y_{ij} and Y_{ji} can be removed in our framework.

5.1.2 Framework Formulation

We propose to prune outliers globally. To this end, we introduce an unknown variable γ_{ij} for each element of Y which indicates whether Y_{ij} is an outlier. We thus aim to estimate both γ_{ij} for outlier detection and β for the relative attribute value prediction in a unified framework. Specifically, for each edge $i \to j \in E$, Y_{ij} is modelled as,

$$Y_{ij} = \beta^T \phi_i - \beta^T \phi_j + \gamma_{ij} + \varepsilon_{ij} \tag{5.1}$$

where *Gaussian* noise $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ with the variance σ; $\gamma_{ij} \in R$ is a sparse symmetric outlier variable which has higher magnitude than σ. When $\gamma_{ij} \neq 0$, Y_{ij} is taken as an outlier. For an edge $i \to j$, if Y_{ij} is not an outlier, we expect $\beta^T \phi_i - \beta^T \phi_j$ should be approximately equal to Y_{ij}, therefore we have $\gamma_{ij} = 0$. On the contrary, when the prediction of $\beta^T \phi_i - \beta^T \phi_j$ differs greatly from Y_{ij}, we can explain Y_{ij} as an outlier and compensate for the discrepancy between the prediction and the annotation with a nonzero value of γ_{ij}. The only prior knowledge we have on γ_{ij} is that it is a sparse variable, i.e. in most cases $\gamma_{ij} = 0$.

For the whole training set, Eq (5.1) is written in its matrix form

$$Y = C\Phi\beta + \Gamma + \varepsilon \tag{5.2}$$

where $Y = [Y_{ij}]$, $\Gamma = [\gamma_{ij}]$, $\varepsilon = [\varepsilon_{ij}]$ and C is the incident matrix of the directed graph G, where $C_{ie} = -1/1$ if the edge e leaves/enters vertex i.

In order to estimate the $I^2 + d$ unknown parameters (I^2 for Γ and d for β), we aim to minimise the discrepancy between the annotation Y and our prediction $C\Phi\beta + \Gamma$, as well as keeping the outlier estimation Γ sparse. To that end, we put a l_2-loss on the discrepancy and a l_1- penalty on the outlier variables as a regularisation measure. This gives us the following cost function:

$$\min_{\beta,\Gamma} \frac{1}{2}\|Y - C\Phi\beta - \Gamma\|_2^2 + \lambda\|\Gamma\|_1 \tag{5.3}$$

$$:= \sum_{i \to j \in E} \left[\frac{1}{2}(Y_{ij} - \gamma_{ij} - \beta^T\phi_i + \beta^T\phi_j)^2 + \lambda|\gamma_{ij}|\right] \tag{5.4}$$

where λ is a free parameter corresponding to the weight for the regularisation term. With this cost function, our Unified Robust Learning to Rank (URLR) framework identifies outliers globally by integrating all local pairwise comparison together.

To solve Eq (5.3), we rewrite the cost function as,

$$L(\beta,\Gamma) = \frac{1}{2}\|Y - X\beta - \Gamma\|_2^2 + \lambda\|\Gamma\|_1. \tag{5.5}$$

where $X = C\Phi$. With $\frac{\partial L}{\partial \beta} = 0$, we have

$$\hat{\beta} = (X^TX)^\dagger X^T(Y - \Gamma). \tag{5.6}$$

The Moore-Penrose pseudo-inverse of X^TX is equivalent to the limit of ridge regression solution: $(X^TX)^\dagger = \lim_{\mu \to 0}((X^TX)^T \cdot (X^TX) + \mu\mathbf{1})^{-1}(X^TX)^T$, where $\mathbf{1}$ is the identity matrix. To avoid numerical instability in many practical applications, we can replace the pseudo-inverse with ridge regression by setting $\mu > 0$. The standard solvers for Eq (5.6) will require $O(I^3)$ computational complexity. To reduce the complexity, the Krylov iterative and algebraic multi-grid methods [HKW10] can be used.

Now plugging the solution of $\hat{\beta}$ back into Eq (5.5) and defining the hat matrix $H = H(X) = X(X^TX)^{-1}X^T$, we have

$$\hat{\Gamma} = \arg\min_{\Gamma} \|Y - \Gamma - H(Y - \Gamma)\|_2^2 + \lambda \parallel \Gamma \parallel_1 \tag{5.7}$$

The first term in Eq (5.7) is L_2- loss of the residuals of the observations $Y - \Gamma$ without the outliers Γ which is: $r = Y - \Gamma - H(Y - \Gamma) = (I - H)(Y - \Gamma)$. Eq (5.12) is thus simplified into into intervening sub-problems: outlier detection in (5.7) and estimation of β using (5.6). And Eq (5.7) does not rely on the estimation of $\hat{\beta}$. Note that for large-scale dataset with large-size

X matrix, it usually means very high computational cost for the hat matrix H. Thus, for outlier detection part in (5.7), we can further simplify Eq (5.7) by Singular Value Decomposition (SVD),

$$X = U\Sigma A^T \tag{5.8}$$

where $U = [U_1, U_2]$ with U_1 an orthogonal basis of the column space of X (i.e. $\mathrm{im}(\mathrm{X})$) and U_2 the orthogonal basis of the kernel space of X (i.e. $\mathrm{Ker}\left(X^T\right)$)[1]. So we have $U_2^T U_2 = I$ and $U_2 X = 0$ (i.e. $U_2 H = 0$). A is the conjugate transpose of U. So the first part in Eq (5.7) can be simplified as,

$$\begin{aligned}
&\|Y - \Gamma - H(Y-\Gamma)\|_2^2 \\
&\quad = \quad (Y - \Gamma - H(Y-\Gamma))^T U_2^T U_2 (Y - \Gamma - H(Y-\Gamma)) \\
&\quad = \quad (U_2(Y - \Gamma - H(Y-\Gamma))^T (U_2(Y - \Gamma - H(Y-\Gamma)) \\
&\quad = \quad (U_2 Y - U_2\Gamma)^T (U_2 Y - U_2\Gamma) \\
&\quad = \quad \|U_2^T Y - U_2^T \Gamma\|_2^2
\end{aligned}$$

Eq (5.7) is now a standard Least Absolute Shrinkage and Selection Operator (LASSO) estimator [FL01],

$$\hat{\Gamma} \quad = \quad \arg\min_{\Gamma} \|U_2^T Y - U_2^T \Gamma\|_2^2 + \lambda \|\Gamma\|_1 \tag{5.9}$$

5.1.3 The Advantage of URLR Over Majority Voting

Figure 5.1(a) illustrates why our URLR framework is advantageous over the local majority voting method for outlier detection. Assume there are five images $A - E$ with five pairs compared three time each, and the correct ranking order of these 5 images in terms of interestingness is $A < B < C < D < E$. Figure 5.1(a) shows that among the five compared pairs, majority voting can successfully identify four outlier cases: $A > B$, $B > C$, $C > D$, and $D > E$, but not the fifth one $E < A$. However when considered globally, it is clear that $E < A$ is an outlier because if we have $A < B < C < D < E$, we can deduce $A < E$. Our formulation can detect this tricky outlier. More specifically, if the estimated β makes $\beta^T \phi_A - \beta^T \phi_E > 0$, it has a small local inconsistency cost for that minority vote edge $A \to E$. However, such β value will be 'propagated' to other images by using the voting edges $B \to A$, $C \to B$, $D \to C$, and $E \to D$, which are accumulated into much bigger global inconsistency with the annotation. This makes our model detect $E \to A$

[1]Note that SVD is one of the most common ways to solve the orthogonal basis U_2 of kernel space $\mathrm{Ker}\left(X^T\right)$.

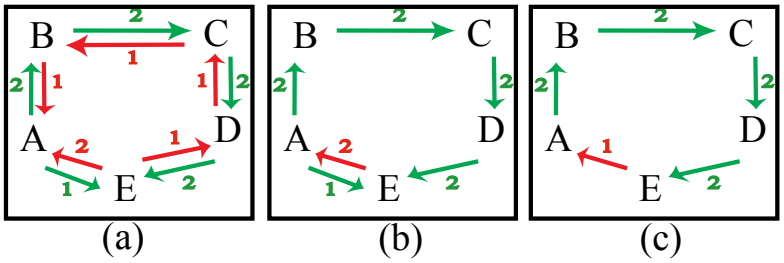


Figure 5.1: Better outlier detection can be achieved using our URLR framework than majority voting. Green arrows indicate correct annotations, while red arrows are outliers.

as an outlier, contrary to the majority voting decision. In particular, the majority voting will introduce a loop comparison $A < B < C < D < E < A$ which is the well-known Condorcet's paradox [Geh83]. We further give two more extreme cases in Fig. 5.1(b) and (c). Due to such Condorcet's paradox, in Fig. 5.1(b) the estimated β from majority voting is even worse than that from all annotation pairs which at least save the right annotation $A \to E$. Furthermore, Fig. 5.1(c) shows that when each pair only receives votes along one direction, majority voting will cease to work altogether, but our URLR can still detect outliers by examining the global cost.

5.1.4 Connection to Robust Ranking

Our formulation can be taken as one special case of Huber-Lasso[2]. We discuss the connections in this section. Huber [Hub81] proposed the following robust ranking (in regression form) with Huber's loss function

$$\min_{\beta} \sum_{i,j} \rho_{\lambda}(Y - C\Phi\beta) \tag{5.10}$$

$$= \min_{\beta} \sum_{i,j} \rho_{\lambda}(\beta^T(\phi_i - \phi_j) - Y_{ij}) \tag{5.11}$$

where the Huber's loss function $\rho_{\lambda}(x)$ is defined as

$$\rho_{\lambda}(x) = \begin{cases} x^2/2, & \text{if } |x| \leq \lambda \\ \lambda|x| - \lambda^2/2, & \text{if } |x| > \lambda. \end{cases}$$

When $|\beta^T(\phi_i - \phi_j) - Y_{ij}| < \lambda$, the comparison is taken as a 'good' one and penalized by l_2-loss for Gaussian noise. Otherwise, it is regarded as the sparse outlier and penalized by

[2]Usually, Huber-Lasso is only for ranking problems, while ours is a learning to rank problem.

l_1-loss. The Huber M-estimator is equivalence to Huber-Lasso [Gan07] in Eq (5.3). Specifically, given the annotation Y of the training data, Huber-LASSO estimates the global ranking order θ by

$$\begin{aligned} \hat{\theta} = & \quad \min_{\theta} \quad \frac{1}{2}\|Y - C\theta - \Gamma\|_2^2 + \lambda \parallel \Gamma \parallel_1 \\ := & \quad \sum_{(i,j)\in E} \quad \left[\frac{1}{2}(Y_{ij} - \gamma_{ij} - \theta_i + \theta_j)^2 + \lambda|\gamma_{ij}|\right] \end{aligned} \tag{5.12}$$

where θ_i is the ranking score for instance i. Eq (5.12) is equivalent to the robust regression problem with Huber's loss function [Hub81], and is called Huber-LASSO.

Our URLR model is extended from Huber-LASSO for the ability of predicting the relative attribute values. It introduces the prediction model parameter β estimated as $\hat{\beta} = \Phi^{\dagger}\hat{\theta}$, where $\dagger$ indicates the moore-penrose pseudo inverse. But this is not the most critical differences – one could still use Huber-LASSO to remove outliers and then use the same Eq (5.12) to estimate β. The more important difference is that URLR can better identify outliers, especially for sparse graphs. More specifically, to solve Eq (5.12), a similar formulation as Eq (5.9) can be used, solved by the same regularisation path method as in URLR. However, instead of SVD decomposing X in Eq (5.8), for Huber-LASSO, the matrix C is decomposed. This means the solution space of Eq (5.12) is $\dim(\Gamma) = |E| - I + 1$ where $|E|$ is the number of pairs compared and I is the number of graph nodes, i.e. training images or videos. Given a sparse dataset, this space is very small. In contrast, URLR enlarges $\dim(\Gamma)$ by including the subspace of the original node space orthogonal to the feature space (Eq (5.9)). This means the solution space of Eq (5.9) is $\dim(\Gamma) \approx |E| - d$. When the feature dimension d is smaller than the number of images/videos I, the dimension of the solution space of Γ for URLR is higher than that of Huber-LASSO, leading to better outlier detection capability. Typically, we have $d < I$ in a large dataset; however if not, it can be made so by reducing the feature dimension.

5.2 Solution of URLR by Regularisation Path

5.2.1 Problem Decomposition and Outlier Detection by Regularisation Path

Note that tuning the regularisation parameter λ in Eq (5.9) is notoriously difficult [SO11, Hub81, FL01]. Especially in our URLR framework, the λ value directly decides the ratio of outliers detected and the ratio is unknown. A number of methods for determining λ exist, but none is suitable for our formulation: (1) some heuristics rules like $\lambda = 2.5\hat{\sigma}$ are popular in existing

robust ranking models such as the M-estimator [Hub81][3]. However setting a constant λ value independent of dataset is far from optimal because the ratio of outliers may vary for different crowdsourcing experiments. (2) Cross validation is also not applicable here because each edge $i \to j$ is associated with a γ_{ij} variable and any held-out edge $i \to j$ also corresponds to an unknown variable γ_{ij}. As a result, cross validation can only optimise part of the sparse variables while leaving those for the held-out validation set undetermined. (3) The other alternatives e.g. Akaike information criterion (AIC) and Bayesian information criterion (BIC) employ the relative quality and likelihood functions of the statistical models as the criterion for parameter selections. These statistical criteria however have no direct connection to the outliers pruned. Ideally λ should be a data-dependent parameter which selects a cut-off value and corresponds to the pruning rate p as the portion of the outliers among all comparisons.

This inspires us to sequentially consider all available solutions for all sparse variables along the Regularisation Path (RP) by gradually decreasing the value of the regularisation parameter λ from ∞ to 0. Specifically, based on the piecewise-linearity property of LASSO [FL01], RP can be efficiently computed by Least Angle Regression (LARS [EHJT04]). When $\lambda = \infty$, the regularisation parameter will strongly penalise outlier detection: if any annotation is taken as an outlier, it will greatly increase the value of the object function in Eq (5.9). When λ is changed from ∞ to 0, LASSO[4] will first select the variable subset accounting for the highest variances to the observations $U_2^T Y$ in Eq (5.9). These high variances should be assigned higher priority to represent the nonzero elements[5] of Γ of Eq (5.2), because Γ compensates the discrepancy between annotation and prediction. Based on this idea, we can order the edge set E by the λ values according to which nonzero γ_{ij} appears first when λ is decreased from ∞ to 0. In other words, if an edge γ_{ij} becomes nonzero at a larger λ_{ij} value, it has a higher probability to be an outlier. Following this order, we identify the top $p\%$ edge set Λ_p as the annotation outliers. And its complementary set $\Lambda_{1-p} = E \setminus \Lambda_p$ are the inliers. Therefore, the outcome of estimating Γ using Eq (5.9) is a binary outlier indication matrix $F_\Gamma = \left[F_{\gamma_{ij}}\right]$:

$$F_{\gamma_{ij}} = \begin{cases} 1 & i \to j \in \Lambda_{1-p} \\ 0 & i \to j \in \Lambda_p \end{cases}$$

[3] $\hat{\sigma}$ is a Gaussian variance and is manually set by human prior knowledge.

[4] For a thorough discussion from a statistical perspective, please read [FL01, FTS12, EHJT04, SO11].

[5] This is related with LASSO for covariate selection in a graph. Please read [MB06] for more details.

Algorithm 1 Learning a unified robust learning to rank model.

Input: A training dataset Φ with pairwise annotation Y and an outlier pruning rate $p\%$.

Output: Detection of outliers F_Γ and prediction model parameter β.

1. Perform SVD on X using Eq (5.8);
2. Solve Eq (5.9) using Regularisation Path;
3. Take the top $p\%$ pairs as outliers and estimate the outlier indicator matrix F_Γ;
4. Compute β using Eq (5.13).

where each element $F_{\gamma_{ij}}$ indicates whether the corresponding edge $i \to j$ is an outlier or not. With this matrix, β can be solved by

$$\beta = (X^T X)^\dagger X^T (Y \odot F_\Gamma) \tag{5.13}$$

where $\odot$ is the Hardmard product and $F_\Gamma = \left[F_{\gamma_{ij}}\right]$. The pseudo-code of learning our URLR model is shown in Alg. 1. Note that it is very efficient to solve the entire regularisation path by LARS: "roughly the same computational cost as a single least square fit" (Pg.438 by Murphy[Mur12]).

5.3 Experiments

5.3.1 Experiment Settings

Datasets We conduct two set of experiments. The first set of experiments are used to statistically validate the efficacy of our framework. We firstly design a statistical simulated experiment to compare different methods of solving Eq (5.9). Then we further validate that our framework can beat the other alternatives on different graphs sparsity, and outlier ratio. For this purpose, we employ the FG-NET image age dataset [FGH10] which contains 1002 images of 82 individuals labeled with age 0 to 69. The training set is composed of the images of 41 randomly selected people and the rest as testing. All experiments are repeated through 10 rounds of training/testing split to reduce variability. Each image is represented by a 55 dimension vector extracted by active appearance models (AAM).

The second set of experiments are conducted on several relative attribute datasets: two image and video interestingness datasets and two general image relative attribute datasets. These datasets are summarised in Table 5.1. The image interestingness dataset consists of 2222 images, each

Dataset	No. pairs	No. img/video	Feat. Dim.	No. cls
Image Age [FGH10]	–	1002	55	–
Image Int.[IXTO11]	16000	2222	932(150)	1
Video Int. [JYF+13]	60000	420	1000(60)	14
PubFig [KBBN09, KPG12]	2616	772	557(100)	8
Scene [OT01, KPG12]	1378	2688	512(100)	8

Table 5.1: Dataset summary for relative attributes. We use the original features to learn the ranking model in Eq (5.13) and reduce the feature dimension (values in brackets) using KPCA to improve outlier detection in Eq (5.9) by enlarging the solution space.

represented as a 932 dimensional feature vector as in [GGR+13]. 16000 pairwise comparisons were collected by [GGR+13] using AMT and are used as annotation.

The video interestingness dataset is the YouTube interestingness dataset introduced in [JYF+13], which contains 14 different categories, each of which has 30 YouTube videos. $10 \sim 15$ annotators were asked to give complete interesting comparisons for all the videos in each category. So the original annotation is noisy but not sparse. We use a bag-of-words of Scale Invariant Feature Transform (SIFT) and Mel-Frequency Cepstral Coefficient (MFCC) as the feature representation which are shown to be effective in [JYF+13] for predicting video interestingness.

We also carry out experiments on two relative attributes datasets – PubFig [KBBN09] and Scene [OT01] to test our URLR model's ability to predict other more general relative visual attributes. PubFig and Scene considered 11 ('smiling', 'round face', etc.) and 6 ('openness', 'natural' etc.) relative attributes respectively. Pairwise attribute annotation was collected by AMT [KPG12]. Each pair was annotated by 5 crowdsourced workers. Gist and colour histograms features are used for PubFig, and Gist alone for Scene. Each image also belongs to a class (celebrity or scene type). These two datasets were designed for classification, with attribute scores as the representation, so the classification accuracy is determined by the attribute prediction accuracy. The detailed of all these dataset are in Chapter 2.1.6.

Evaluation metrics Since image age ranking experiments validate our framework from a statistical aspect, we use Kendall tau rank correlation to measure the statistical association between predicted rankings and ground-truth rankings[6]. Higher Kendall tau correlation means higher

[6]Recent statistical theories [RA14, JLYY11] show that the dense human annotations collected in [GGR+13] and [JYF+13] can give a reasonable approximation of ground truth for interestingness.

correlations of two ranking orders. While for the image and video interestingness dataset, we prefer to use the metric – Kendall tau rank distance[7] to measure the rank correlation between the predicted ranking order and the ground truth ranking of unseen test data provided by [GGR+13] and [JYF+13] respectively. Higher Kendall tau rank distance means lower quality of the ranking order predicted. The similarities and differences between rank distance and rank correlation are discussed in [Car09]. We use both of them to evaluate different experiments in our settings. For the scene and pubfig image dataset, the relative attributes are very sparsely collected and their prediction performance can only be evaluated indirectly by image classification accuracy with the predicted relative attributes as image representation.

Competitors We compare our method (URLR) with four competitors.

Jiang *et al.* [JYF+13] this method uses majority voting for outlier pruning and rankSVM for learning to rank.

Gygli *et al.* [GGR+13] this method also first removes outliers by majority voting. After that, the fraction of selections by the pairwise comparisons for each data point is used as an absolute interestingness score and a regression model is then learned for prediction.

***Huber-LASSO* [XXHY13, FTS12]** this is a statistical ranking method that performs outlier detection as described in Sec. 5.1.4, followed by estimating β by $\hat{\beta} = \Phi^{\dagger}\hat{\theta}$. (4)

Raw this is our URLR model without outlier detection, that is, all annotations are used to estimate β.

5.3.2 Learning to Rank Image Age

We design this image age experiment to further validate the statistical significance of our URLR framework over the alternatives on graph sparsity and outlier ratio.

5.3.2.1 Crowdsourcing errors.

We use the ground truth age to generate the pairwise comparisons without any error (used for Real+RR). Errors are synthesized according to human error patterns estimated by data collected

[7] The Kendall tau ranking distance between two lists $L1$ and $L2$ is

$$K(\tau_1, \tau_2) = \mid \{(i,j) : i < j, (\tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j)) \vee (\tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j))\} \mid$$

, where $\tau_1(i)$ and $\tau_2(i)$ are the rankings of the element i in $L1$ and $L2$.

by an online pilot study [Fu]: 4000 pairwise image comparison from 20 skilled willingly participating "good" workers are collected as *unintentional errors.* The human unintentional age error pattern is built by fitting the error rate against true age difference between collected pairs. As expected, humans are more error-prone for smaller age difference. We use this error pattern for unintentional error generations. *Intentional errors* are introduced by 'bad' workers who insert a random order. This is easily simulated by adding random comparisons. In practice, human errors in crowdsourcing experiments can be a mixture of both types. Thus two settings are considered: *Unint.*: errors are generated following the estimated human unintentional error model resulting in around 10% errors. *Unint.+Int.*: random comparisons are added on top of *Unint.*, giving an error ratio of around 25%, unless otherwise stated. Note that these percentage of errors are only used for the age experiments. Since the ground-truth of testing data is known to us, we can give an upper bound for all the algorithms by using the ground-truth as the training data which is – *GT*.

5.3.2.2 *Quantitative results.*

Four experiments are performed using different settings to show the effectiveness of our method quantitatively.

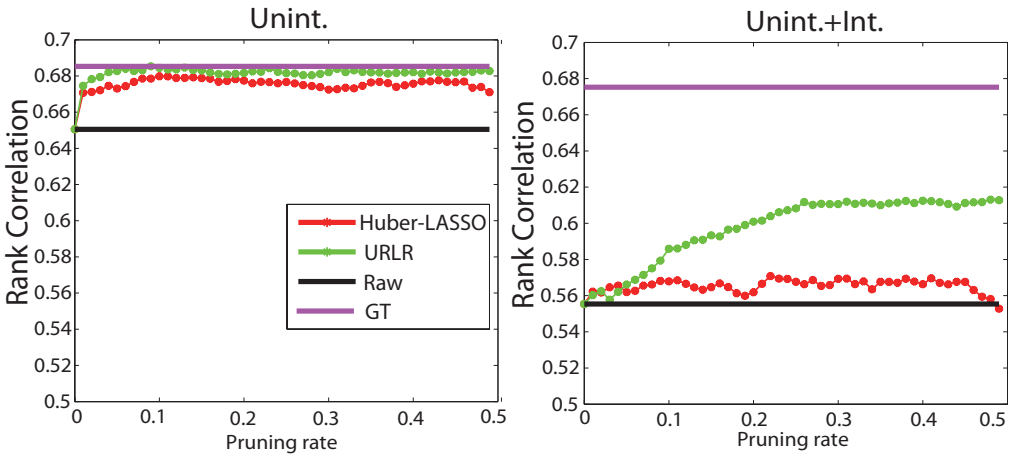


Figure 5.2: Comparing *URLR* and *Huber-LASSO* on ranking prediction under two error settings.

1. *Effectiveness of our models.* For 300 training images, 600 unique comparisons are sampled. Fig. 5.2 shows that *URLR* and Huber-LASSO improve over *Raw* indicating that outliers are effectively pruned. Both models are robust to low error rate (Fig. 5.2 Left: 10% in Unint.), whilst the performance of *URLR* is significant better than *Huber-LASSO* in high error ratio (Fig. 5.2 Right: 25% in Unint.+Int.) because of using low-level feature

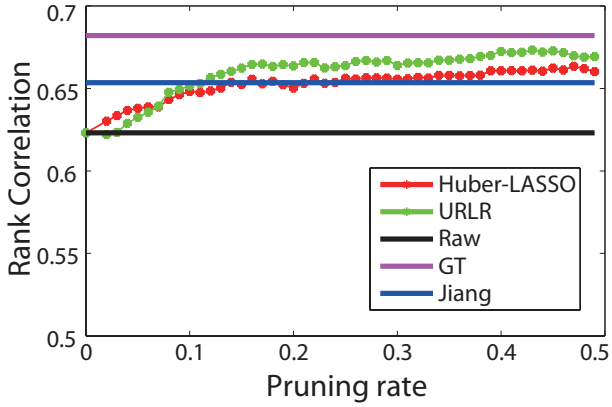


Figure 5.3: Comparing *URLR* and *Huber-LASSO* against *Jiang et al.* [JYF+13].

representation to increase the dimension of model inconsistency subspace Γ and $\dim(\Gamma)$ from 301 for *Huber-LASSO* to 546 for *URLR*. This result validates our analysis that higher $\dim(\Gamma)$ leads to better chance of identify more accurate outliers.

2. *Comparison with Jiang's majority voting.* Given the same data but each pair compared by 5 workers under the Unint.+Int. error condition, Fig. 5.3 shows that *Jiang* beats *Raw*. This shows that for relative dense graph, majority voting is still a good strategy of removing some outliers and improves the prediction accuracy. However, *URLR* outperforms *Jiang et al.* [JYF+13] after the pruning rate passes 10%. This demonstrates that aggregating all paired comparisons globally for outlier pruning is more effective than aggregating them locally for each edge as done by majority voting.

3. *Effects of graph sparsity.* For 300 training images, one comparison per edge is sampled for $400-4000$ edges under the Unint.+Int. error setting. Fig. 5.5 shows that the when the graph becomes very sparse, the ability of *Huber-LASSO* to detect outliers diminishes because the dimension of the model inconsistency space, $\dim(\Gamma)$ decreases with the number of edges $|E|$. In contrast, *URLR*'s performance decreases much more gracefully due to the use of low-level features. Particularly, *URLR* remains very effective yielding an AUC of 0.85 given 400 edges for 300 nodes with 25% error ratio– an extremely noisy sparse graph, whilst *Huber-LASSO* gives an AUC value around half of that and a ranking prediction performance identical to that of *Raw*. The gap between *URLR* and *Raw* suggests that our model becomes more useful given more sparse graphs.

4. *Effects of error ratio.* We use the Unint.+Int. error model to vary the amount of ran-

dom comparisons and simulate different amounts of errors in 10 sampled graphs from 300 training images and 2000 unique sampled pairs. The pruning rate is fixed at 25%. Fig. 5.4 shows that *URLR* remains effective even when error ratio reaches as high as 35%. This shows that although a sparse outlier model is assumed, our model can deal with non-sparse outliers.

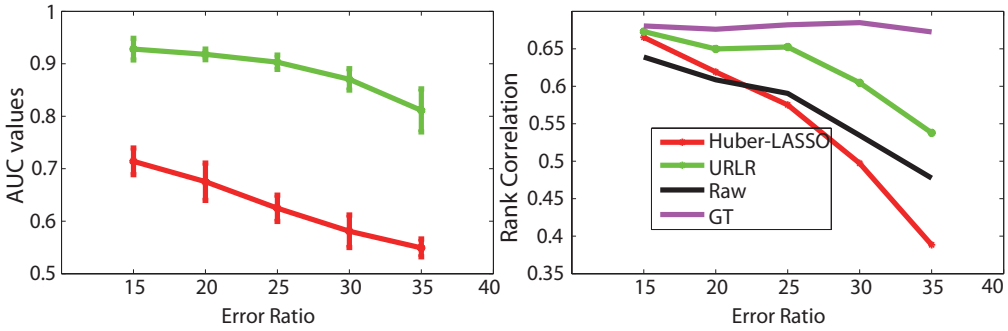


Figure 5.4: Effects of error ratio.

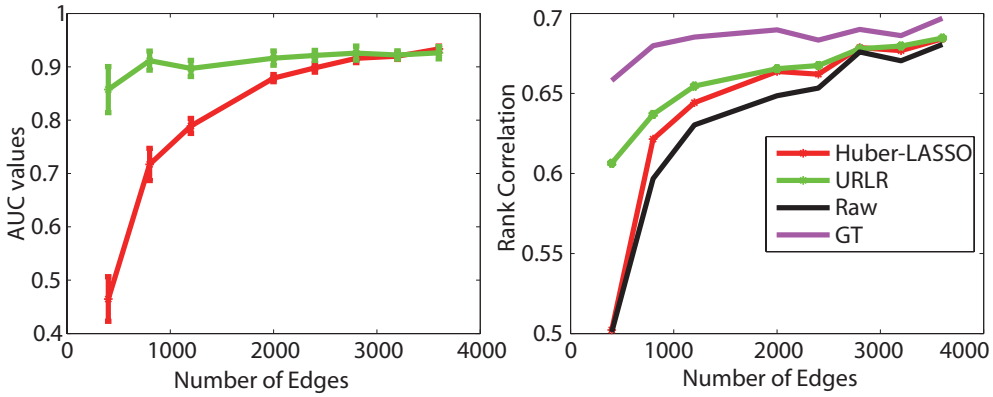


Figure 5.5: Effects of graph sparsity.

5.3.2.3 *Qualitive results.*

1. **What are pruned and in what Order?** The regularisation path can be examined as λ decreases to produce a ranked list for all pairwise comparisons according to outlier probability. Figure 5.6 shows the relationship between the pruning order (i.e. which pair is pruned first) and ground truth age difference, illustrated by examples. It is seen that overall outliers with larger age difference tend to be pruned first. This means even with a conservative pruning rate, obvious outliers (potentially causing more performance degradation in learning) can be reliably pruned by our model.

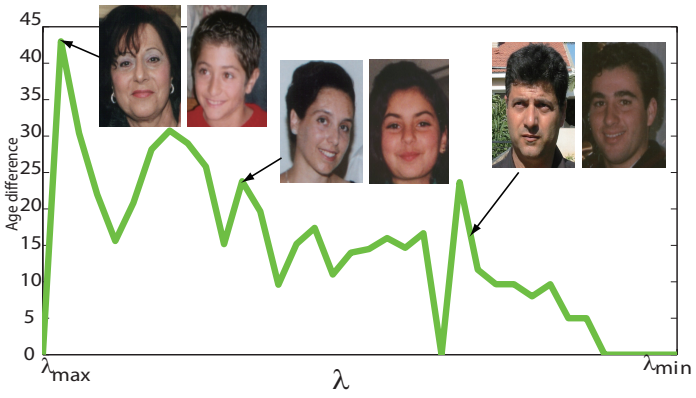


Figure 5.6: Relationship between the pruning order and actual age difference for RHRL+.

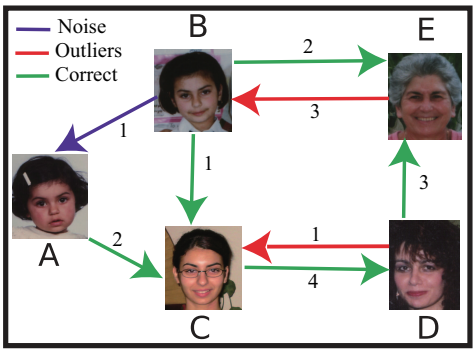


Figure 5.7: Left: the crowdsourced pairwise labels with the number of votes. Colour indicates different types of pairs. Right: the outlier probabilities of each pair via our model.

2. **What went wrong for majority voting.** To intuitively explain the advantage of our model over majority voting, we carry out a small-scale experiment using five images (A,B,C,D,E) with 17 comparisons by multiple workers (see Fig.5.7). The outlier $D \to C$ (i.e. D is older than C) can be effectively dealt with by majority voting as more correctly voting. However, the opposite happens for $E \to B$ and majority voting also fails flat for the single noisy label of $B \to A$. These three cases naturally reflect the success, failure and no-effect cases for majority voting in dealing with crowdsourcing noise. More subtly but critically, in this example majority voting will induce Condorcet's paradox in that we have $B \to C \to D \to E \to B$ which means that the error in $B \to E$ has propagated across the graph without any global inconsistency check. In contrast, global inconsistency among pairs is modeled via Hogerank decomposition in our method to identify outliers even if they receive a majority vote. In particular, Fig. 5.7 shows that $E \to B$ has the the highest outlier probability; it will thus be removed by our model to prevent the error propagation.

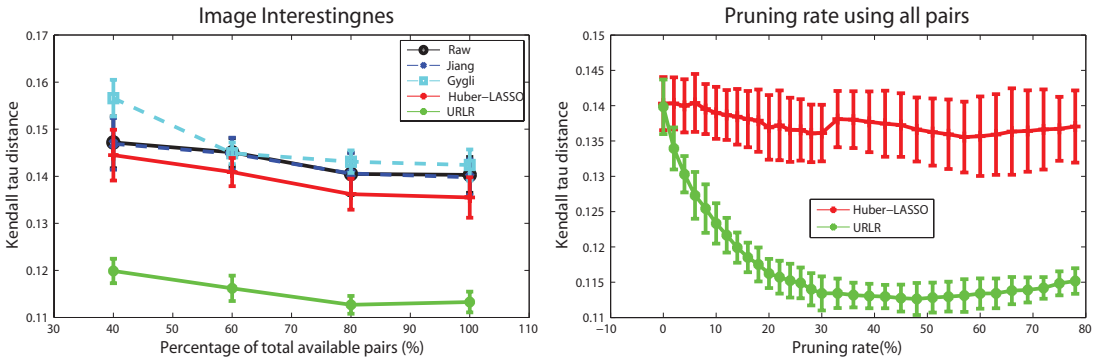


Figure 5.8: Image interestingness prediction performance. Lower is better.

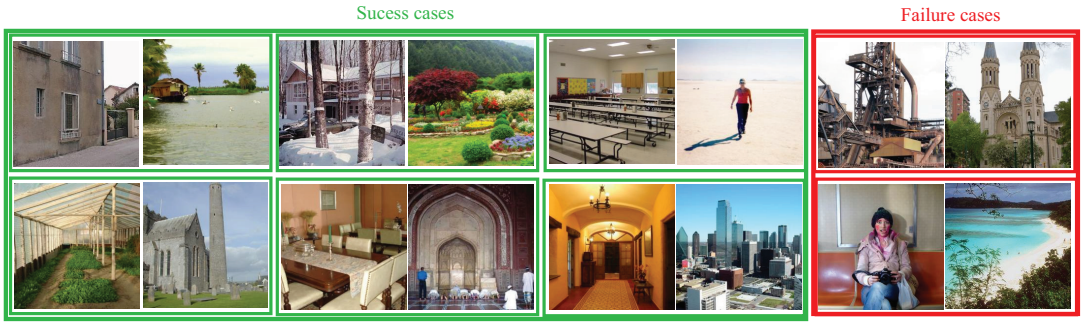


Figure 5.9: Qualitative results of interestingness prediction. The left images were annotated as more interesting than the right ones. The success cases show the true positive outliers detected by URLR. Two failure cases are shown in red boxes where inliers are incorrectly removed.

5.3.3 Image Interestingness Prediction

5.3.3.1 Experimental settings

For this experiment, we randomly select 1000 images for training and the remaining 1222 are used for testing. All the experiments are repeated 10 times to reduce variance. The pruning rate p is set to 20%. We also vary the number of annotated pairs used to test how well each compared method copes with increasing annotation sparsity.

5.3.3.2 Comparative results

The results are shown in Fig. 5.8 (a). It shows clearly that our *URLR* significantly outperforms the four alternatives for a wide range of annotation density. This validates the effectiveness of our method. In particular, the improvement over Jiang *et al.* [JYF+13] and Gygli *et al.* [GGR+13] demonstrates the superior outlier detection ability of *URLR*. *URLR* is superior to *Huber-LASSO* because the joint outlier detection and ranking estimation framework of *URLR* enables the enlargement of the solution space of Eq (5.9), resulting in better outlier detection performance. The

performance of *Gygli et al.* [GGR+13] is the worst among all methods compared, particularly so given sparser annotation. This is not surprising – in order to get an reliable absolute interestingness value, dozens or even hundreds of comparisons per image are required, a condition not met by this dataset. The estimated value becomes less reliable given sparser annotations, explaining the worse relative performance. The performance of *Huber-LASSO* is also better than *Jiang et al.* [JYF+13] and *Gygli et al.* suggesting even a weaker global outlier detection approach is better then the majority voting based local one. Interestingly even the baseline method Raw gives a comparable result to *Jiang et al.* [JYF+13] and *Gygli et al.* [GGR+13] which suggests that just using all annotations without discrimination in a global cost function Eq (5.5) is as effective as majority voting.

Fig. 5.8 (b) evaluates how the performances of *URLR* and *Huber-LASSO* are affected by the pruning rate p. It can be seen that the performance of *URLR* is improving with an increasing pruning rate. This means that our *URLR* can keep on detecting true positive outliers. The gap between *URLR* and *Huber-LASSO* gets bigger when more comparisons are pruned showing *Huber-LASSO* stops detecting outliers much earlier on. Some qualitative results of outlier detection using *URLR* are shown in Fig. 5.9.

5.3.4 Video Interestingness prediction

5.3.4.1 Experimental settings

Because comparing videos across different categories is not very meaningful, we follow the same settings as in [JYF+13] and only compare the interestingness of videos within the same category. Specifically, we use 20 videos and their paired comparison for training and the remaining 10 videos for testing. The experiments are repeated for 10 rounds and the averaged results are reported. We use rankSVM with χ^2 kernel which is approximated by additive kernel of explicit feature mapping [VZ11]. Kendall tau rank distance is used, and we find that the same results are obtained if the prediction accuracy used in [JYF+13] is used instead. The pruning rate is again set to 20%.

5.3.4.2 Comparative results

The results of video interestingness prediction are shown in Fig 5.10. Fig. 5.10(a) compares different methods given varying amounts of annotations, and Fig. 5.10(b) shows the per category performance. The results show that all the observations we had for the image interestingness pre-

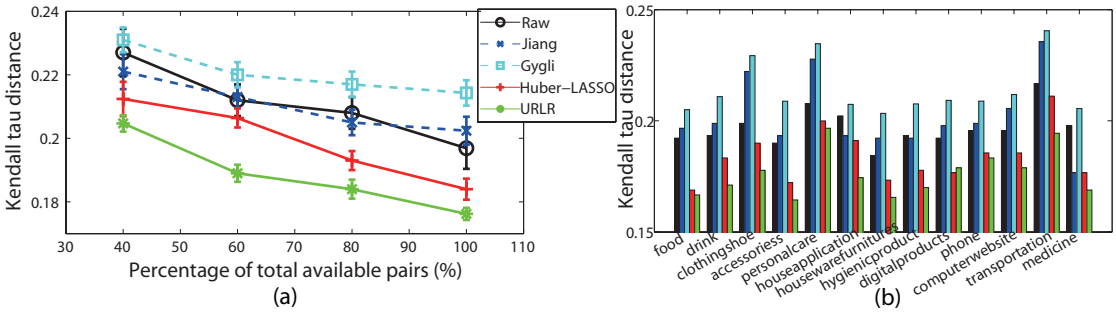


Figure 5.10: Video interestingness prediction results.

diction experiment still hold here, and across all categories. However in general the gaps between our *URLR* and the alternatives are smaller as this dataset is densely annotated. In particular the performance of *Huber-LASSO* is much closer to our *URLR* now. This is because the advantage of *URLR* over *Huber-LASSO* is stronger when $|E|$ is close to I. Given a dense pairwise annotation $|E|$ is much greater than I and the effect of enlarging the solution space diminishes.

5.3.5 Relative Attributes Prediction for Image Classification

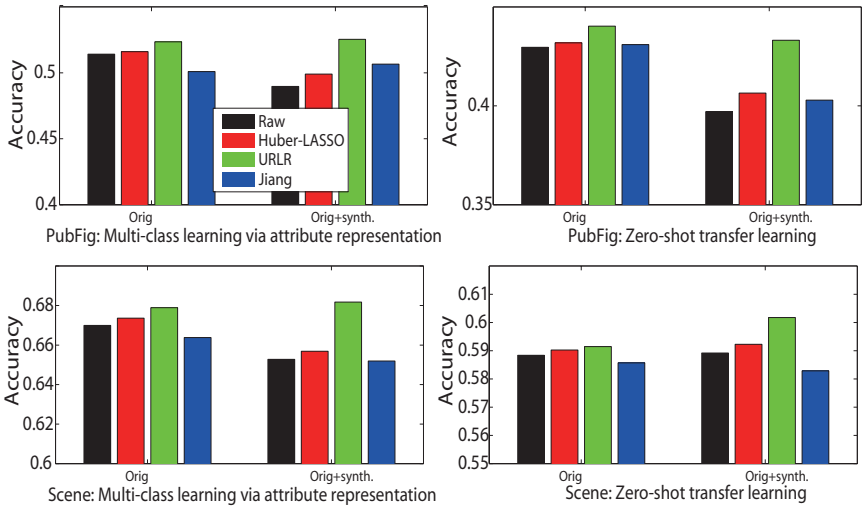


Figure 5.11: Relative attribute performance evaluated indirectly as image classification rate (chance = 0.125).

5.3.5.1 Experimental settings

We evaluate image classification with relative attributes as representation on the PubFig and Scene datasets under two settings: multi-class classification where samples from all classes are

available for training and zero-shot transfer learning where one class is held out during training (a different class is used in each trial with the result averaged). Our experiment setting is similar to that in [PG11b], except that image-level, rather than class-level pairwise comparisons are used. Two variations of the setting are used:

- *Orig:* This is the original setting with the pairwise annotations used as they are.
- *Orig+synth*: By visual inspection, there are limited annotation outliers in these datasets, perhaps because the relative attributes are less subjective compared to interestingness. To simulate more challenging situations, we randomly add 150 random comparison for each attribute, many of which would correspond to outliers. This will lead to around 20% extra outliers.

The pruning rate is set to 7% for original dataset (*Orig*) and 27% for dataset with additional outliers inserted for all attributes of both datasets (*Orig+synth*).

5.3.5.2 Comparative results

Without the ground truth of relative attribute values, different models are evaluated indirectly via image classification accuracy in Fig. 5.11. Note that the method of *Gygli et al.* [GGR$^+$13] is not compared here as the annotation is too sparse for it to learn a meaningful model. The following observations can be made:

1. Our URLR always outperforms *Huber-LASSO*, *maj-voting* (Jiang) and *Raw* for all experiment settings. The improvement is more significant when the data contain more errors (*Orig+synth*).
2. The performance of other methods is in general consistent to what we observed in the image and video interestingness experiments: *Huber-LASSO* is better than majority voting (Jiang *et al.* [JYF$^+$13]) and Raw often gives better results than majority voting too.
3. It is noted that for PubFig, *Jiang et al.* [JYF$^+$13] is better than Raw given more outliers, but it is not the case for Scene. This is probably because the annotators are very familiar with the celebrity faces in PubFig.. Consequently there should be more subjective/intentional errors for Scene, causing majority voting to choose wrong local ranking orders (e.g. not many people are sure how to compare the relative values of the 'diagonal plane' attribute for two images). These majority voting + outlier cases can only be rectified by using a global approach such as our *URLR*, even the *Huber-LASSO* method to a certain extent.

Figure 5.12: Qualitative results on image relative attribute prediction.

5.3.5.3 *Qualitative Results*

Figure 5.12 gives some examples of the pruned pairs for both datasets using URLR. In the success cases, the left images were (incorrectly) annotated to have more of the attribute than the right ones. However, they are either wrong or too ambiguous to give consistent answers, and as such are detrimental to learning to rank. A number of failure cases (false positive pairs identified by *URLR*) are also shown. Some of them are caused by unique view point (e.g. Hugh Laurie's mouth is not visible, so it is hard to tell who smiles more; the building and the street scene are too zoomed in compared to most other samples); others are caused by the weak feature representation, e.g. in the 'male' attribute example, the colour and Gist features are not discriminative enough for judging which of the two men has a more 'male' attribute.

5.4 Summary

We have proposed a novel unified robust learning to rank (URLR) framework for predicting image and video interestingness. The key advantage of our method over the existing majority voting based approaches is that we can detect outliers globally by minimising a global ranking inconsistency cost. The joint outlier detection and ranking estimation formulation also provides our model with an advantage over the conventional statistical ranking methods such as Huber-LASSO for outlier detection. The effectiveness of our model in comparison with state-of-the-art alternatives has been validated using image and video interestingness datasets. Further, it is generally applicable to other relative attribute prediction tasks as demonstrated by our relative attribute based image classification experiments.

Chapter 6

Conclusions and Future Work

This thesis explored the problem of attribute learning for image and video understanding. In particular, we

1. studied learning latent attributes for understanding complex image and video data with very sparse, incomplete and ambiguous annotations of user-defined attributes in Chapter 3;

2. solved the projection domain shift, prototype sparsity and the inability to combine multiple semantic representation problems by transductive multi-view embedding;

3. finally we investigated the robust learning of relative attributes from crowdsourced pairwise comparisons.

It is clear that the work in this thesis is unable to cover all the potential useful applications and generalisation of attribute learning for image and video understanding. Other directions such as attribute classifier from linguistic descriptions [ESE13, PG11a] and image retrieval by attribute feedback [PP12, SFD11] are also promising directions for attribute learning. Nevertheless, we believe that our three problems touch the challenging topics and are of significant contributions to the fields of computer vision and machine learning in general. Also, we believe that the research on attribute learning is just at the beginning and our efforts make the problem – attribute learning for image/video understanding one step closer to the *Holy Grail* of *life-long learning* in visual recognition for the computer vision and machine learning communities.

6.1 Learning Latent Attributes

Learning multi-modal latent attributes is studied in Chapter 3. We introduce a semi-latent attribute space, which enables the use of sparse, incomplete and ambiguous prior annotated knowledge available from both user-defined and two types of automatically discovered latent attributes. By formulating a computationally tractable solution via a novel and scalable topic model, we show latent attributes computed by our framework can be utilised to tackle a wide variety of learning tasks in the context of multimedia content understanding including multi-task, label-noise, N-shot and surprisingly zero-shot learning.

Nevertheless, there remain a number of important open questions to be addressed. Thus far, our attribute-learner has not yet considered inter-attribute correlation explicitly [QHR$^+$07, THW$^+$09], though this limitation is shared by most other attribute learners with the exception of [LKS11]. For our task, this can be addressed straightforwardly by generalizing the correlated topic model (CTM) [BL07] instead of the conventional LDA [BNJ03], which should produce commensurate gains in performance to those observed elsewhere [LKS11].

The complexity of our model in terms of the size of the attribute/topic-space was fixed to a reasonable value throughout, and we focused on learning with attribute-constraints on the topics. A more desirable solution would be a non-parametric framework which could infer the appropriate dimension of the latent attribute-space automatically given available UD attributes.

6.2 Transductive Multi-view Embedding

Intrinsically the attribute learning framework belongs to the scope of *learning to learn* or *life-long learning* [PY10, TM95, PL14] which studies how to intelligently apply previously learned knowledge to perform well on future recognition tasks. It is thus possible to understanding our transductive multi-view embedding framework from the perspective of machine learning.

For the zero-shot learning problem, this thesis in Chapter 4 has presented a transductive multi-view embedding framework that not only rectifies the projection shift and prototype sparsity problems, but also exploits the complementary of multiple semantic representations of visual data. Such a framework enables the TMV-HLP algorithms to greatly improve both zero-shot and N-shot learning tasks as well as a number of novel cross-view annotation tasks. Extensive experiments are carried out and the results show that our approach significantly outperforms existing methods for both zero-shot and N-shot recognition on three image and video benchmark datasets.

With regard to the projection domain shift and prototype sparsity problems, still a number of problems have not been solved and remain as future work.

1. In this thesis multi-view Canonical Component Analysis (CCA) is employed for learning the embedding space. Although it works well, other embedding framework can be considered (e.g. Wang *et al.* [WHW+13]). In particular, in the current pipeline of Chapter 4, low-level features are firstly projected onto different semantic views before the views are embedded. Since the projection itself can be considered as an embedding, it is possible to develop a unified embedding framework to combine these two embedding steps together.

2. Although the presented framework is designed for lifelong learning [PL14] and under a realistic lifelong learning setting [CSG13], an unlabelled data point could either belong to a seen class (those in the auxiliary/source dataset) or an unseen class. The current framework needs to be extended to firstly distinguish these two types of data before performing zero-shot recognition.

3. Our results suggest that more views, either manually defined (attributes), extracted from linguistic knowledge bases (word space), or learned directly from visual data (deep learning features), give rise to better embedding space. More investigations are needed to enable more systematic design and selection of semantic views for embedding.

I should also be noted that our framework can explore the correlations of labels on a zero-shot learning problem and solve the problem of multi-label zero-shot learning. As an extension of our transductive multi-view embedding framework, we developed the multi-label zero-shot learning framework in [FYH+14b], and proposed two tailor-made multi-label algorithms – DMP and TraMP. The experimental results on benchmark multi-label datasets show the efficacy of our framework for multi-label zero shot learning over a variety of baselines. Besides the proposed tailor-made multi-label algorithms – DMP and TraMP, our strategy in [FYH+14b] could potentially help to generalise the existing multi-label algorithms to solve multi-label zero-shot learning problems. Thus it would be extremely interesting to investigate these in the future.

6.3 Robust Learning of Relative Attributes

In this thesis, we propose a novel robust approach learning for relative attributes from noisy and sparse pairwise comparison data. In particularly, our framework can tackle the problems

of detecting outliers and estimating ranking scores jointly in our unified framework. In Chapter 5, we demonstrate both theoretically and experimentally that our method is superior to existing majority voting based methods as well as statistical ranking based methods.

In Chapter 5, we discuss some connetions of our URLR framework with the Huber-Lasso. We formulate the outlier detection on graphs as a standard LASSO problem, and solve it by regularisation path. As future work, we need to study a better solution for such a LASSO formulation and also investigate other penalty functions [SO11] for outlier detection beside LASSO. For example She *et al*. [SO11] proposed a non-convex hard-threshold penalty function which has been shown better ability for outlier detection than that of the L_1-penalty in LASSO formulation.

References

[APHS13] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[Arr63] Kenneth Arrow. *Social Choice and Individual Values, 2nd Ed.* Yale University Press, New Haven, CT, 1963.

[AWST09] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, 2009.

[BBS10] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010.

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013.

[BDGS05] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 2005.

[Ben09] Yoshua Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, pages 1–127, 2009.

[BETG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BFK09] John Blitzer, Dean P. Foster, and Sham M. Kakade. Zero-shot domain adaptation: A multi-view approach. Technical report, TTI-TR-2009-1, 2009.

[Bie87] Irving Biederman. Recognition by components - a theory of human image understanding. *Psychological Review*, 1987.

[BJ03] David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[BL07] David M. Blei and John Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.

[BM07] David M. Blei and Jon McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BP13] A. Biswas and D. Parikh. Simultaneous active learning of classifiers and attributes via relative feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[BPV+92] Peter F. Brown, Vincent J.Della Pietra, Peter V.deSouza, Jenifer C.Lai, and Robert L.Mercer. Class-based N-gram models of natural language. *Journal Computational Linguistics*, 1992.

[BT78] Harry G. Barrow and J.Martin Tenenbaum. Recovering intrinsic scene characeteristics from images. *Computer Vision Systems*, 1978.

[BZM07a] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *IEEE International Conference on Computer Vision*, 2007.

[BZM07b] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, 2007.

[Car09] Ben Carterette. On rank correlation and the distance between rankings. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2009.

[CB09] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Conference on Empirical Methods on Natural Language Processing*, 2009.

[CB13] Xi Chen and Paul N. Bennett. Pairwise ranking aggregation in a crowdsourced setting. In *ACM International Conference on Web Search and Data Mining*, 2013.

[CECC08] Vitor R. Carvalho, Jonathan L. Elsas, William W. Cohen, and Jaime G. Carbonell. A meta-learning approach for robust rank learning. In *SIGIR 2008 LR4IR - Workshop on Learning to Rank for Information Retrieval*, 2008.

[CGXL13] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Chang Loy. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[CK10] Olivier Chapelle and S. Sathiya Keerthi. Efficient algorithms for ranking with SVMs. *Inf. Retr.*, 2010.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CQL+07] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *International Conference on Machine Learning*, 2007.

[CSG13] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *IEEE International Conference on Computer Vision*, 2013.

[CSVZ14] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[CWCL09a] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourceable qoe evaluation framework for multimedia content. In *ACM MM 2009*, 2009.

[CWCL09b] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. Crowdsourceable QoE evalutation framework for multimedia content. In *ACM International Conference on Multimedia*, 2009.

[Dav88] H. David. *The Methods of Paired Comparisons, 2nd Ed.* Griffin's Statistical Monographs and Courses, 41. Oxford University Press, New York, NY, 1988.

[DDS+09] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[DFV11] Ankur Datta, Rogerio Feris, and Daniel Vaquero. Hierarchical ranking of facial attributes. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2011.

[DJV+14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, 2014.

[DOB11] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[DTXM09] Lixin Duan, Ivor W. Tsang, Dong Xu, and Stephen J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 2004.

[ESE13] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *IEEE International Conference on Computer Vision*, December 2013.

[FBWT10] Rob Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. Semantic label sharing for learning with many categories. In *European Conference on Computer Vision*, 2010.

[FCS+13] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Neural Information Processing Systems*, 2013.

[FEH10] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2352–2359, 2010.

[FEHF09] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[FGH10] Yun Fu, Guodong Guo, and T.S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[FGZ+10] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010.

[FHST13] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision*, 2013.

[FHX+14a] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Zhengyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, 2014.

[FHX+14b] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. Interestingness prediction by robust learning to rank. In *European Conference on Computer Vision*, 2014.

[FHXG12] Yanwei Fu, Timothy Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, 2012.

[FHXG13] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Learning multi-modal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[FHXG14] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot recognition and annotation. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 2001.

[FTS12] Jianqing Fan, Runlong Tang, and Xiaofeng Shi. Partial consistency with sparse incidental parameters. *arXiv:1210.6950*, 2012.

[Fu] Yanwei Fu. Online Pilot Age Study of Pairwise Comparison of Human Face Images. `http://www.eecs.qmul.ac.uk/~yf300/survey4/`.

[FYH+14a] Yanwei Fu, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-calss and multi-label zero-shot learning. In *European Conference on Computer Vision'14 workshop on Parts and Attribute*, 2014.

[FYH+14b] Yanwei Fu, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-label zero-shot learning. In *British Machine Vision Conference*, 2014.

[FZ07] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Neural Information Processing Systems*, December 2007.

[Gan07] Iréne Gannaz. Robust estimation and wavelet thresholding in partial linear models. *Stat. Comput.*, 17:293–310, 2007.

[Geh83] William V. Gehrlein. Condorcet's paradox. *Theory and Decision*, 1983.

[GGR+13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *IEEE International Conference on Computer Vision*, 2013.

[GKIL13] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 2013.

[GY14] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *International Conference on Machine Learning*, 2014.

[HG11] Sung Ju Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, 2011.

[HGX11a] Tim Hospedales, Shaogang Gong, and Tao Xiang. Learning tags from unsegmented videos of multiple human actions. In *International Conference on Data Mining*, 2011.

[HGX11b] Timothy Hospedales, Shaogang Gong, and Tao Xiang. Video behaviour mining using a dynamic topic model. *International Journal of Computer Vision*, 2011.

[HKW10] Anil N. Hirani, Kaushik Kalyanaraman, and Seth Watts. Least squares ranking on graphs. *arXiv:1011.1716*, 2010.

[HLGX11] Tim Hospedales, Jian Li, Shaogang Gong, and Tao Xiang. Identifying rare and subtle behaviours: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[HLM09] Yuchi Huang, Qingshan Liu, and Dimitris Metaxas. Video object segmentation by hypergraph cut. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[HLZM10] Yuchi Huang, Qingshan Liu, Shaoting Zhang, and Dimitris Metaxas. Image retrieval via probabilistic hypergraph ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[Hot36] Harold Hotelling. Relations between two sets of variables. *Biometrika*, 1936.

[HSG11] Sung Ju Hwang, Fei Sha, and Kristen Grauman. Sharing features between objects and their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[HSMN12] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Association for Computational Linguistics 2012 Conference*, 2012.

[HSST04] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. In *Neural Computation*, 2004.

[Hub81] P. J. Huber. *Robust Statistics*. New York: Wiley, 1981.

[HYL+07] A. Hauptmann, Rong Yan, Wei-Hao Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958 –966, aug. 2007.

[HYLC13] Chaoqun Hong, Jun Yu, Jonathan Li, and Xuhui Chen. Multi-view hypergraph learning by patch alignment framework. *Neurocomputing*, 2013.

[IPTO11] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Neural Information Processing Systems*, 2011.

[IXTO11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[JLYY11] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Math. Program.*, 2011.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[JWZ13] Jungseock Joo, Shuo Wang, and Song-Chun Zhu. Human attribute recognition by rich appearance dictionary. In *IEEE International Conference on Computer Vision*, 2013.

[JYC+11] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval*, 2011.

[JYF+13] Yu-Gang Jiang, YanranWang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and predicting interestingness of videos. In *AAAI Conference on Artificial Intelligence*, 2013.

[JYN10] Yu-Gang Jiang, Jun Yang, and Chong-Wah Ngo. Representations of keypoint-based semantic concept detection: a comprehensive study. *IEEE Transaction on Multimedia*, 2010.

[KBBN09] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[KCS08] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *ACM Computer-Human Interaction (CHI) Conference on Human Factors in Computing Systems*, 2008.

[KJYFF11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[KKTH12] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[KPG12] Adriana Kovashka, Devi Parikh, and Kristen Grauman. WhittleSearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012.

[Lam09] Christoph H. Lampert. Kernel methods in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2009.

[Lap05] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, September 2005.

[LEB08] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*, 2008.

[LGL+08] Yuting Liu, Bin Gao, Tie-Yan Liu, Ying Zhang, Zhiming Ma, Shuyuan He, and Hang Li. Browserank: letting web users vote for page importance. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2008.

[LHK13] Chengjiang Long, Gang Hua, and Ashish Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *IEEE International Conference on Computer Vision*, 2013.

[LHS+13] Xi Li, Weiming Hu, Chunhua Shen, Anthony Dick, and Zhongfei Zhang. Context-aware hypergraph construction for robust spectral clustering. *IEEE Transactions on Data and Knowledge Engineering*, 2013.

[LJLFF10] Yongwhan Lim Li-Jia Li, Hao Su and Li Fei-Fei. Objects as attributes for scene classification. In *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, Crete, Greece, September 2010.

[LKS11] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[LLS+13] Xi Li, Yao Li, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Contextual hypergraph modelling for salient object detection. *IEEE International Conference on Computer Vision*, 2013.

[LNH09] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[LNH13] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2004.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 2006.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representation in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[Mil95] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[MSN11] Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In *IEEE International Conference on Computer Vision*, pages 1227–1234, 2011.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[MWN+09] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Conference on Empirical Methods on Natural Language Processing*, 2009.

[MYP11] Michae Maire, Stella X. Yu, and Pietro Perona. Object detection and segmentation from joint embedding of parts and pixels. In *IEEE International Conference on Computer Vision*, 2011.

[NAS09] David Newman, Arthur Asuncion, and Padhraic Smyth. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

[NWFF08] Juan C. Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[OT01] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: Aholistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 2001.

[PG11a] Devi Parikh and Kristen Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[PG11b] Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, 2011.

[PH12] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[PHPM09] Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems*, 2009.

[PL14] Anastasia Pentina and Christoph H. Lampert. A PAC-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

[PP12] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *European Conference on Computer Vision*, 2012.

[PY10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Data and Knowledge Engineering*, 22(10):1345–1359, 2010.

[QHR+07] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In *ACM International Conference on Multimedia*, 2007.

[RA14] Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, 2014.

[RES13] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *Neural Information Processing Systems*, 2013.

[RSS+10] Marcus Rohrbach, Michael Stark, Gyorgy Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where – and why? semantic relatedness for knowledge trans-

fer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 910–917, 2010.

[RSS12] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[SEZ+14] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2014.

[SF08] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[SFD11] Behjat Siddiquie, Rogerio Feris, and Larry Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[SFF10] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[SGS+13] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Neural Information Processing Systems*, 2013.

[SHH+07] Cees G. M. Snoek, Bouke Huurnink, Laura Hollink, Maarten de Rijke, Guus Schreiber, and Marcel Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9:975–986, 2007.

[SI07] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[SJY08] Liang Sun, Shuiwang Ji, and Jieping Ye. Hypergraph spectral learning for multi-label classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[SO11] Yiyuan She and Art B. Owen. Outlier detection using nonconvex penalized regression. *Journal of American Statistical Association*, 2011.

[SOJN08] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods on Natural Language Processing*, 2008.

[SQTW09] Zhengya Sun, Tao Qin, Qing Tao, and Jue Wang. Robust sparse rank learning for non-smooth ranking measures. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, 2009.

[SS07] Amos J Storkey and Masashi Sugiyama. Mixture regression for covariate shift. In *Neural Information Processing Systems*, 2007.

[SSG12] Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Constrained semi-supervised learning via attributes and comparative attributes. In *European Conference on Computer Vision*, 2012.

[STT11] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[SZ03] Josef Sivic and Andrew Zisserman. Video google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[TAP+10] George Toderici, Hrishikesh Aradhye, Marius Pasca, Luciano Sbaiz, and Jay Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3447–3454, 2010.

[Thr96] S. Thrun. *Learning To Learn: Introduction*. Kluwer Academic Publishers, 1996.

[THW+09] Jinhui Tang, Xian-Sheng Hua, Meng Wang, Zhiwei Gu, Guo-Jun Qi, and Xiuqing Wu. Correlative linear neighborhood propagation for video annotation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(2):409–416, 2009.

[TM95] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 1995.

[TYH+09] Jinhui Tang, Shuicheng Yan, Richang Hong, Guo-Jun Qi, and Tat-Seng Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM International Conference on Multimedia*, 2009.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 2008.

[vdSGS08] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[vdSGS10] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[VZ11] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[Wat04] Duncan J. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. University Presses of California, 2004.

[WBL09] Chong Wang, David M. Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[WBU10] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 2010.

[WBW+11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[WCCL13] Chen-Chi Wu, Kuan-Ta Chen, Yu-Chun Chang, and Chin-Laung Lei. Crowdsourcing multimedia qoe evaluation: A trusted framework. *IEEE TMM*, 2013.

[WG07] Yong Wang and Shaogang Gong. Translating topics to words for image annotation. In *ACM International Conference on Conference on Information and Knowledge Management*, 2007.

[WHG11] Ou Wu, Weiming Hu, and Jun Gao. Learning to rank under multiple annotators. In *International Joint Conference on Artificial Intelligence*, 2011.

[WHW+13] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision*, 2013.

[WJ13] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. *IEEE International Conference on Computer Vision*, 2013.

[WM09] Yang Wang and Greg Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774, 2009.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.

[WZ11] Xianwang Wang and Tong Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM International Conference on Multimedia*, 2011.

[XHJ+12] Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 2012.

[XHY12] Qianqian Xu, Qingming Huang, and Yuan Yao. Online crowdsourcing subjective image quality assessment. In *ACM International Conference on Multimedia*, 2012.

[XXHY13] Qianqian Xu, Jiechao Xiong, Qingming Huang, and Yuan Yao. Robust evaluation for quality of experience in crowdsourcing. In *ACM International Conference on Multimedia*, 2013.

[YA10] Xiaodong Yu and Yiannis Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *European Conference on Computer Vision*, 2010.

[YCF+13] Felix X. Yu, Liangliang Cao, Rogerio S. Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[YHWZ11] Kuiyuan Yang, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. Tag tagging: Towards more descriptive keywords of image content. *IEEE Transactions on Multimedia*, 13:662 –673, 2011.

[YP97] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, 1997.

[Yu12] Stella X. Yu. Angular embedding: A robust quadratic criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[ZB07] Dengyong Zhou and Christopher J. C. Burges. Spectral clustering and transductive learning with multiple views. *International Conference on Machine Learning*, 2007.

[ZBL+04] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Neural Information Processing Systems*, pages 321–328. MIT Press, 2004.

[ZHS06] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: clustering, classification, and embedding. In *Neural Information Processing Systems*, 2006.

[Zhu07] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison Department of Computer Science, 2007.

[ZMWH07] Zheng-Jun Zha, Tao Mei, Zengfu Wang, and Xian-Sheng Hua. Building a comprehensive ontology to refine video concept detection. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pages 227–236, New York, NY, USA, 2007. ACM.

[ZWX+13] Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, and Shuang Liu. Robust relative attributes for human action recognition. *Pattern Analysis and Applications*, 2013.